Using Protein Microarrays to Profile Human PTMs

by

Johnathan P Neiswinger

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March, 2014

Abstract

Over the last quarter century, protein microarray technology has emerged as a prominent field in scientific study. The versatility of the platform, coupled with the ability to characterize thousands of proteins in a parallel and high-throughput manner, has resulted in great strides in our knowledge database.

Many clinical studies have used protein microarrays as analytical tools to identify biomarkers using analytical protein microarrays. These allow for the detection of varying expression levels of proteins in a cell lysate as well as binding affinities and specificities of a sample. One newer technique that has emerged uses cell or tissue lysates as the arrayed substance (reverse-phase protein microarray). This powerful tool allows for the determination of altered protein modifications from two different samples, allowing researchers to identify potential biomarkers in diseased tissue. Perhaps the most well-used microarray type is the functional protein microarray, or "target protein microarray." The ability to performed hundreds or thousands of individual reactions in parallel becomes an unbiased and powerful tool that scientists can use to draw conclusions at a global or systematic level.

As science and technology continues to advance, new approaches to old paradigms are often challenged. My thesis projects involve utilizing protein microarrays and bioinformatics to change how we view phosphorylation events and the cross-talk between phosphorylation and glycosylation.

The wealth of information in databases continues to grow as larger proteomic-wide studies are carried out and deposited for all to use. This allows bioinformaticians to predict new interactions that were never possible before. One such interaction involves

scaffolding proteins, proteins that can interact with at least two other proteins in signaling pathways. Scaffolding proteins have been studied for years and have been found to play critical roles in cellular signal transduction. In this study, human protein-protein interaction (PPI) and kinase-substrate relationship (KSR) networks were used to predict scaffolding proteins involved in phosphorylation signal transduction. We predicted 212 scaffolding proteins involving 612 non-redundant phosphorylation pathways. One third (359 of the 1,103 known KSRs) of the phosphorylation-mediated signaling pathways are known to be regulated by at least one scaffold protein.

We examined that the predicted scaffolding proteins are enriched for protein domains known to interact with phosphorylation sites, and exhibit similar characteristics of other known scaffolding proteins. Intriguingly, the predicted scaffolding proteins tend to have large protein sizes, perhaps due to their ability to adapt to multiple interactions with other proteins. Furthermore, these proteins are more evolutionarily conserved, suggesting important roles in different biological processes across species. When comparing to other human proteins, scaffolding proteins also contain more known phosphorylation sites, indicating that the scaffolds themselves might be regulated by phosphorylation-mediated signal transduction.

In order to test these predictions, microarrays were employed on the human proteome microarray, which contains over 17,000 full-length human proteins. *CSNK2A1* (with predicted scaffolds *PIN1* and *ATF2*) and *MAPK9* (with predicted scaffold *ATF2*) were used to treat these microarrays in the presence of ^{33}P-γ-ATP. After careful alignment and scoring, the predicted scaffolding proteins *PIN1* and *ATF2* were shown to

mediate 28 distinct phosphorylation events. Through this initial study, we have shown that our initial predictions appear to hold merit.

Mutations and dysregulation of kinases play causal roles in human disease, development, cell signaling, and metabolism. Understanding the function of kinases continues to be of interest for biomarker discovery as well as for the development of agonists and antagonists for use in disease therapy. O-linked β-N-acetylglucosamine (*O*-GlcNAc) is a post-translational modification known to regulate a variety of protein functions, including localization, enzyme activity, and protein stability. Like phosphorylation, *O*-GlcNAcylation modifies serine and threonine residues on nuclear and cytoplasmic proteins, and is a ubiquitous and reversible process that regulates cellular signaling. Recent evidence indicates that site-specific crosstalk between *O*-GlcNAcylation and phosphorylation as well as the *O*-GlcNAcylation of kinases plays an important role in regulating cell signaling. Therefore, it is very important to study the *O*-GlcNAcylation of the kinome. Previous studies utilizing a functional kinase array were able to identify 42 kinases as substrates of O-Glycosyltransferase (OGT) using an *in vitro* OGT assay with [H^3] radiolabeling. While some promising results were obtained, the limitations of [H^3] labeling demand a more sensitive approach if one is to examine a larger library of proteins.

Herein, using multiple *in vitro* OGT labeling assays and immunofluorescent detection techniques, we were able to obtain a high confidence hit list based on shared hits from three different detection methods. A kinome array was fabricated, which contains 350 unique full-length human kinases representing approximately 70% of the human kinome. After *in vitro* labeling by OGT, arrays were treated either with one of two

Cy5-labeled antibodies that recognize *O*-GlcNAcylated residues (RL2 and CTD110.6) or with β-1,4-galactosyltransferase (GalT1 (Y289L)), which transfers azido-modified galactose (GalNAz) from UDP-GalNAz to *O*-GlcNAc residues of modified proteins. A simple click chemistry reaction followed with incubation of Alexa Fluor® 647 DIBO alkyne to fluorescently label the hits. Data analysis revealed that these hits were very reproducible and a total of 104 kinases were shared between all three methods.

Many hits were subsequently validated both *in vitro* and *in vivo*, further validating our methods. Among the hits, *O*-GlcNAcylated sites were identified via mass spectrometry for *BRSK2* and *PAK4*, with mutagenesis studies on the latter validating the identification. Through this simple, yet sensitive strategy, we have shown with high confidence that at least 20% of the human kinome (and 31% of those kinases tested) is glycosylated and the dataset created will likely spawn many further validation studies in the future.

In conclusion, we show the utility of both bioinformatics and microarrays to predict novel functions of proteins and to probe an entire family of proteins for post-translational modifications.

Thesis Advisors:

Heng Zhu, PhD

Jin Zhang, PhD

Readers:

Heng Zhu, PhD

Jin Zhang, PhD

Joel Pomerantz, PhD

Thesis Committee:

Heng Zhu, PhD

Jin Zhang, PhD

Joel Pomerantz, PhD

Gerald Hart, PhD

Jiang Qian, PhD

Acknowledgements

I would like to thank my mentors, Drs. Heng Zhu and Jin Zhang for their mentorship and understanding as I completed my thesis work. Their wealth of knowledge and passion for new technology were critical in steering me through my various projects, while allowing for freedom to determine the proper path for completion. Their invaluable input in experimental designs, to technological input, to manuscript preparations were critical to my success.

I would like to give special thanks to my committee members and thesis readers Drs. Joel Pomerantz, Gerald Hart, and Jian Qiang for their support and insight in the direction to go on my projects.

I would like to thank Dr. Robert Newman and Dr. Hee Sool Rho for their valuable mentorship and guidance in learning the intricacies of protein microarrays and protein purification. A big thanks to Dr. Jianfei Hu as well for spearheading the scaffolding project and providing me with useful analysis for all my various projects throughout the years. I would also like to thank Dr. Guanghui Han for collaborating with me on the O-glycosylation project and for the many discussions regarding the state of the O-GlcNAcylation field. I also thank everyone in both the Zhu and Zhang labs for their support and all the fun we have had over the last several years.

Finally, I would like to thank my wife, Lauren, and my four children, Colby, Ella, Aiden, and Ana for their unending love, patience, and understanding as we all completed this journey together. The many long days and nights away from home were not easy to deal with and without their support I would have never been able to finish. I would thank

my parents, Mike and Cindy, for raising me in a loving home with my sister Jacque and brother Ben, nurturing my mind at a young age, and teaching me that all things are possible in Christ. I would like to thank my in-laws Don and Sharyn for helping out with the kids whenever possible, as well as providing love and support to both Lauren and me as we balanced life on a graduate students' stipend.

"May the words of my mouth, and the meditation of my heart, be acceptable in Your sight, O Lord, my Strength and my Redeemer" Psalm 19:14

Table of Contents

List of Tables

List of Figures

Chapter 1: Protein Microarrays and Their Applications

1.1 History of Protein Microarrays

In 1983, Tse-Wan Chang spotted a variety of antibodies onto a glass slide in a matrix formation in order to study "the potential of simultaneous multiple determinations of specific cell surface antigens in one reaction incubation (Chang, 1983)." This study represents one of the first concepts of what we today refer to as a microarray. This notion blossomed into ambient analyte immunoassays, introduced by Roger Ekins in 1989, in which two fluorescently labeled antibodies were used to detect the occupation of a "sensing" antibody spotted on a slide (Ekins, 1989). Over the next ten years, the concept was transformed into DNA microarrays, which allow for the detection of mRNA expression levels of thousands of genes in parallel.

However, it was found that mRNA levels do not always correlate with protein expression, so the technology expanded further through the creation of protein microarrays in order to overcome some of these limitations (Gygi et al., 1999; Zhu et al., 2001; Kopf et al., 2007). Although DNA microarrays provide a useful platform for biological studies, they cannot emulate the function of what many consider the workhorses of the cell – proteins.

Protein array technology was developed in the early stages by organizing bacterial strains of a cDNA library on a nylon membrane, lysing the bacteria, and immobilizing the total protein complement (Bussow et al. 1998). Although the technology was promising, it suffered from four critical issues. First, the human proteins were expressed in bacteria,

so they lacked the critical post-translational modifications found in eukaryotes; second, the bacterial lysis was performed in denaturing conditions, resulting in misfolded proteins; third, the proteins of interest were not separated from the thousands of bacterial proteins from the cell lysate; and finally, the overall density of the array was very low. The purity and density problems were quickly resolved by a variety of groups, who used either purified proteins (MacBeath et al., 2000; Zhu et al., 2000; Ge, 2000) or antibodies (Schweitzer et al., 2000) for their arrays.

In 2001, the first study using these high-density arrays was presented by the Snyder group at Yale (Zhu et al., 2001). This "proteome" microarray began by expressing approximately 5,800 full-length yeast open reading frames (ORFs) individually in yeast, where they were purified as N-terminal GST-fusion proteins. The proteins were then spotted on a glass slide in duplicate robotically to form the first high-density "proteome" microarray, as it contained more than 75% of the known yeast proteome. Since this time, many more proteome microarrays have been created from viruses, bacteria, plants, and humans (Zhu et al., 2006; Chen et al., 2008; Popescu et al., 2007; Jeong et al., 2012).

1.2 Types of Protein Microarrays

There are three accepted types of protein microarrays that have been created, all with different functions and benefits: analytical, functional, and reverse-phase. In the next section we will define and compare the differences between these three classes of microarrays (Figure 1.1).

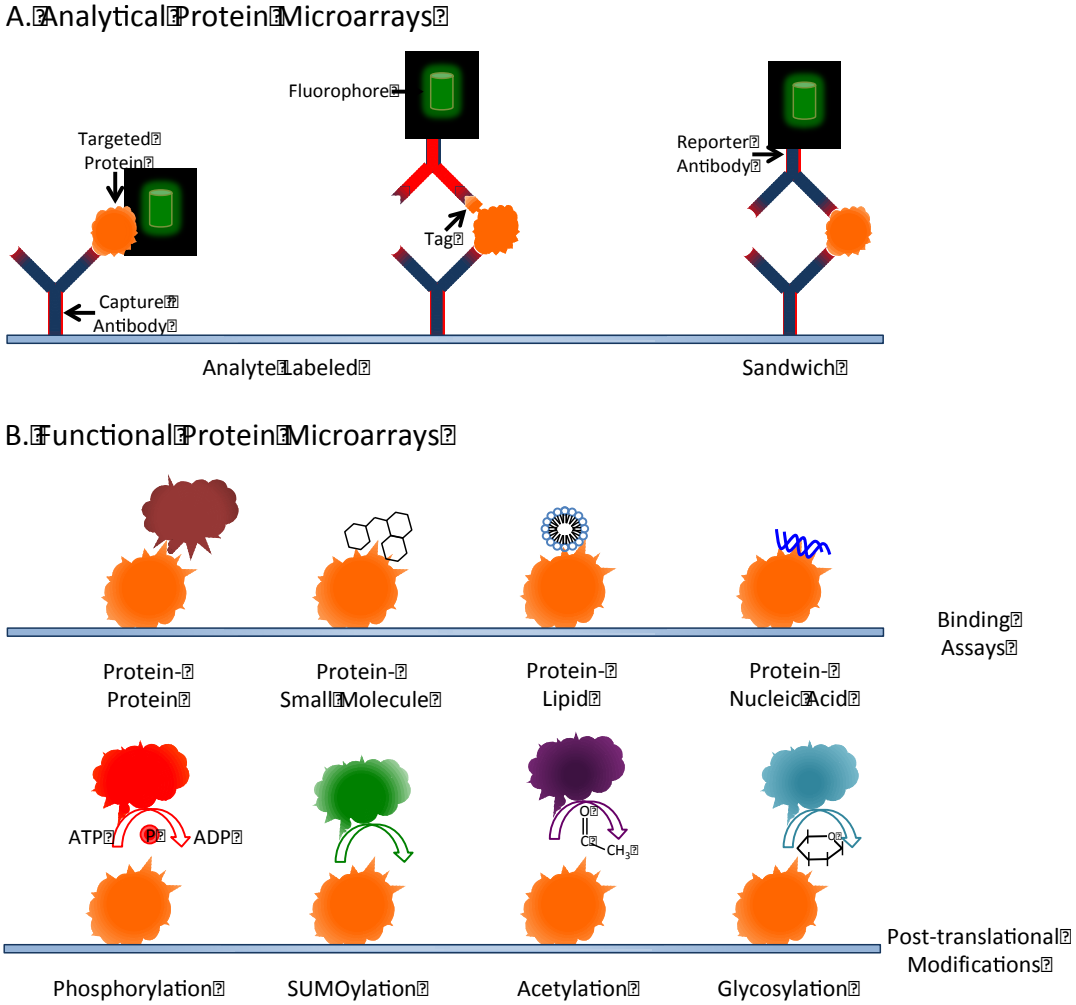

Figure 1.1. The three types of protein microarrays. (A) Analytical protein microarrays generally are fabricated with antibodies on a glass surface. Analyte can be labeled directly, identified via a tag, or identified with a separate labeled antibody. (B) Functional protein microarrays constitute the largest diversity in their applications. Simple binding assays can be performed and are usually tethered to a fluorescent protein, but can also utilize tight non-covalent interactions (i.e. biotin-streptavidin). Post-translational modification studies utilize an enzyme of interest with a labeled co-factor for detections of *de novo* modifications. (C) Reverse-phase protein microarrays consist of probing complex mixtures of proteins, such as a cell lysate, with very specific antibodies. Multicolor assays are possible using this technique, and these are commonly used in biomarker identification studies.

Analytical Protein Microarrays

Analytical microarrays focus on protein detection from a complex mixture of proteins (Figure 1.1 A). The canonical example of this type of array is the antibody array, which uses an assortment of spotted antibodies to capture proteins from a cell lysate. This is most commonly performed using a process called direct protein labeling. In this technique, protein samples are labeled with a fluorophore before binding to the array. Protein labeling can be useful for high-abundance proteins, but suffers from low sensitivity and specificity in target labeling. A sandwich approach was developed to increase the sensitivity by using two separate antibodies to detect the target protein. The "capture" antibody is immobilized onto the solid phase, where it binds the protein of interest. The second "reporter" antibody contains the fluorescent label and binds to the target protein, leading to much greater sensitivity and specificity, even down to femtomolar concentrations (Poetz et al., 2005). Although antibodies are the primary means of protein capture for analytical microarrays, they do have some sizeable disadvantages. For starters, the specificity of the antibodies can vary considerably and can cross-react with many other proteins, resulting in a large amount of false positives. Another hurdle is due to the difficulty in producing large quantities of antibodies in a high-throughput fashion. The advent of recombinant antibodies holds a promising means of overcoming this, but complex production issues need to be addressed before this technology becomes feasible in the research community (Brichta et al., 2005).

Functional Protein Microarrays

The second class of protein microarrays is the functional protein microarray (Figure 1.1 B). These arrays consist of individually purified proteins immobilized to a glass slide and are used in a variety of biochemical reactions or assays. Unlike the analytical arrays, which focus on detections from a complex protein mixture, functional microarrays generally consist of treating the array with a single purified protein in order to study its interaction with the proteins on the chip. These interactions can vary from binding activities (protein-protein, protein-DNA, protein-lipid, protein-drug) to enzyme-substrate relationships, which are analyzed from a biochemical assay performed on the array (phosphorylation, ubiquitylation, acetylation, glycosylation, SUMOylation, nitrosylation) (Poetz et al., 2005; Chen et al., 2006). One large advantage is derived from the sheer amount of diverse experiments that can be applied to the array, but caution must be applied when labeling hits as true *in vivo* interactions. Although many assays involve the use of purified proteins, the interactions may not reflect what actually occurs in a living system. Conversely, some true interactions may not be seen due to the absence of co-factors in the reaction mix. Because of this, any interactions identified on a functional array must be further characterized before conclusions can be drawn.

Reverse-Phase Protein Microarrays

Reverse-phase protein microarrays are unique in that they employ the opposite format to the classical microarray (Figure 1.1 C). Tissues, cell lysates, or cell

fractionations from different states are immobilized to a glass slide, where they can be treated with a variety of probes to identify a protein of interest. Cloud Paweletz was one of the first to use this method in his analysis of histological changes in prostate cancer patients (Paweletz, et al., 2001). In this study, lysates from three different stages of prostate cancer from many patients were immobilized: normal prostate epithelium, prostate intraepithelial neoplasia, and invasive prostate cancer. They were then able to detect and quantify the phosphorylation states of a variety of proteins, including Akt and ERK, throughout the progression of the disease. This technique was sensitive, precise, and linear enough to detect even the smallest changes between samples, and they were able to show a positive correlation between phosphorylation state and prostate cancer progression. This method shows great promise going forward, but like the analytical microarrays, it is limited by the availability and specificity of commercially available antibodies.

Type	Product Name	Company	Type	Protein Content
Protein ID Antibody Arrays	RayBioHuman RTK Phosphorylation Antibody Array	RayBiotech	Analytical	Antibodies against 71 human kinases
	PlasmaScan 380 Antibody Microarray	Arrayit	Analytical	Antibodies for detection of human plasma
	Cytokine Antibody Microarray	Full Moon BioSystems	Analytical	Antibodies against 77 human cytokines
	Kinase Antibody Microarray	Full Moon BioSystems	Analytical	Antibodies against 276 human kinases
Pathway Detection Antibody Arrays	MAPK Pathway Phospho Antibody Array	BioCat	Analytical	185 phospho-protein antibodies for proteins in the MAPK pathways
	Signaling Explorer Antibody Microarray	Full Moon BioSystems	Analytical	1,358 antibodies covering multiple pathways
	Wnt Signaling Phospho Antibody Microarray	Full Moon BioSystems	Analytical	227 phosho-antibodies from cell growh, movement, and development pathways
Human Protein Arrays	ProtoArray	Invitrogen	Functional	9,000 human proteins
	Kinex	Kinexus Bioinformatics	Functional	200 human kinases
Pathogen Arrays	Arrayit Pathogen Antigen Microarray	Arrayit	Functional	Essential proteins of various pathogens
Cell Lysate Arrays	SomaPlex	Protein Biotechnologies	Reverse-Phase	Variety of human cancer cell lysates

Table 1. Commercially Available Protein Microarrays

1.3 Microarray Fabrication

Protein Purification

As mentioned before, one of the largest hurdles in microarray technology has been the high-throughput purification of protein samples. Unlike DNA, which can be synthesized *in situ* with the same properties as in the cell, proteins cannot be synthesized *in vitro* at high efficiency because of their complex chemistry. Therefore, new technologies had to be developed to accommodate the need for very pure and functional proteins.

Recombinant antibodies have provided a useful alternative to the traditional hybridoma-based technology used for high-quality antibody purification. Phage display was one of the earliest techniques used for recombinant antibody production. Using antibody-fragment encoding genes (VH and VL) and bacteriophage capsid gene fusion, large sets of human antibody libraries are stored in prokaryotic systems, where they can be easily expressed by phage infection (Carmen et al., 2002). After expression, they are displayed in the phage capsid and purified using column chromatography. These antibodies also have the advantage of not containing the Fc domain that intact IgG antibodies contain, therefore eliminating non-specific binding to the Fc receptor (Knappik et al., 2009). More recently, other approaches have been developed as well for the production of recombinant antibodies, including the eukaryotic expression system of yeast display (Chao et al., 2006). Using yeast is beneficial in maintaining some of the post-translational modifications that are absent when using bacteria.

The development of functional microarrays faces a large hurdle as well because of the large quantities of highly purified proteins that are needed. Not only do the proteins need to be free of contaminants, they also require conditions that facilitate proper folding, native post translational modifications, and require optimal physical conditions during the purification. To this end, large-scale purifications have been developed using both yeast (*Saccharomyces cerevisiae*) and bacteria (*Escherichia coli*) expressions systems (Zhu et al., 2001; Chen et al., 2008; Jeong et al., 2012; Zhu et al., 2007). Several advantages exist in producing human proteins in yeast rather than in bacteria: (1) higher solubility; (2) higher yields of large proteins (e.g. > 50 kD); (3) better preserved conformation; (4) less immunogenicity; and (5) they contain many, but not all, of the native post-translational modifications that a prokaryotic system cannot provide. While both can utilize batch purification in a 96-well format to purify thousands of proteins at a time, many labs cannot undergo such a large endeavor.

In 2004, Ramachandran et al. developed a novel technique called the nucleic acid programmable protein array (NAPPA). Plasmid DNA is biotinylated and immobilized on a streptavidin surface adjacent to an affinity-tag antibody (e.g. GST). The GST-tagged protein is then synthesized via *in situ* transcription/translation and immediately binds to the antibody. This allows for the generation of over ten thousand proteins on one slide without the tedious cloning and purification steps, and slides can be stored long-term (Ramachandran et al., 2004). However, the array is not "pure" as the proteins are co-localized with the DNA and can provide some unwanted interactions. Alternatively, the PISA method (protein *in situ* array) bypasses the immobilization of DNA on the array by using free DNA in a reaction mixture on an affinity-tagged surface (He et al., 2001).

Angenendt et al. created an array containing over 13,000 spots using a robotic system to accurately mix sub-nanoliter reactions containing the necessary transcription/translation machinery (Angenendt et al., 2006). After the protein is translated in the nanodrops, they bind to the surface of the slide and the DNA is washed away, leaving a "pure" array. While these cell-free methods of creating protein microarrays eliminate the need for large-scale cloning and purifications, they have not been widely utilized because of the low protein yield, the absence of post-translational modifications, and the difficulties in producing large (>60 kDa) proteins.

Surface Chemistry

Once the proteins have been purified, the solid support in which they are immobilized is another factor to be considered. An optimal surface is one that allows for proteins to be tightly bound while still maintaining their three-dimensional conformation, yield high signal-to-noise ratios, and have a long shelf life. Popular types include adsorption, diffusion, covalent immobilization, affinity capture, and metal.

In the early life of this technology, glass slides coated with polyvinylidene fluoride (PVDF), nitrocellulose, or polystyrene were used. Through simple adsorption, proteins were easily arrayed (Stillman et al., 2000). However, because these surfaces are relatively soft, they allowed protein spots to spread out, decreasing the overall density of each spot. It was also found that many applications generate high background for a low signal-to-noise ratio. Another approach uses an agarose/polyacrylamide gel to immobilize proteins in their native conformation and the hydrophilic matrix prevents the

lateral movement found in the PVDF-coated membranes. Because of the restriction in movement of the proteins, the size of each spot is decreased, resulting in a greater number of total proteins allowed on the array (Charles et al., 2004). Aldehyde- or epoxy-derivatized glass surfaces are a third popular surface that forms covalent bonds with the printed proteins. These provide strong binding interactions and relatively low backgrounds, but protein orientation is randomized (MacBeath et al., 2000). If protein orientation is critical, surfaces coated with streptavidin, glutathione, or nickel allow for biotin-, GST-, or His-tagged proteins, respectively, to bind tightly in a non-covalent manner. By having all the proteins in the same conformation, a large increase in signal-to-noise ratio can be achieved (Zhu et al., 2001). The final surface that is currently being used involves the adsorption of proteins onto a surface coated with gold, silver, or steel. In the case of the gold coated arrays, porous gold is deposited onto a bare gold surface and the substrate is patterned with methyl and carboxy-terminated SAMs (self-assembled monolayers). This creates well-defined hydrophilic spots (carboxy-terminated) where the substrates are covalently attached, and a superhydrophobic background is formed (methyl-terminated) that minimizes protein adsorption from biological solutions. Proteins arrayed in this manner facilitate MALDI-TOF detection of protein arrays (Evans-Nguyen et al., 2008). This type of array is also compatible with surface plasmon resonance spectroscopy (SPR) (Huang et al., 2006).

Characterization	Surface	Properties
Adsorption	Polyvinylidene fluoride (PVDF), nitrocellulose, polystyrene	High background signals in certain assays; low cost
Diffusion	Agarose/polyacrylamide gel, hydrogel	Weak protein immobilization; preservation of protein conformation
Covalent Immobilization	Aldehyde, NHS, epoxy, carboxylic ester, etc.	Irreversible protein immobilization; good for covalent reactions; random protein conformation
Affinity Capture	Ni^{2+}-NTA, streptavidin, glutathione	Can control protein orientation
Metal	Gold, silver, steel, etc.	Conductive surface; compatible with SPR, OI-RD, and mass spectrometry detection

Table 2. Summary of microarray surfaces.

Microarray Printing

Careful placement of purified samples is crucial when fabricating microarrays. Spots must be homogenous and dense, yet spatially discrete, therefore robotic systems are often used to "print" a surface with samples. There are two methods of spot formation, contact printing and non-contact printing. Each of these methods can also be done in serial (one sample at a time) or in parallel (multiple samples at a time). In this next section we will discuss the wide variations in these methods as well as their respective advantages and disadvantages (Barbulovic-Nad et al., 2006).

Contact printing involves direct contact between a printing device and the substrate. The printing device first comes in contact with a sample, which adheres to the apparatus via surface tension. It then physically contacts the solid surface, where the sample is attached covalently or non-covalently, depending on which surface is used. These technologies use a variety of tools, including solid pins, split pins, microstamps, and nano-tips. Initial microarrays used a single pin for fabrication, but that was quickly evolved into an array of pins to increase throughput. Pin printing is considered serial printing, while microstamps deposit large amounts of sample in a parallel fashion. Nano-tip printing is based on Atomic Force Microscopy (AFM) and can generate arrays with submicron spots.

Pin printing is one of the most common forms of microarray fabrication. There are a variety of pins that can be used, each with its own characteristics. Solid pins are the most simple and were the first to be used. They consist of a solid piece of metal that is dipped into the sample solution to "load" the tip. Capillary force action allows for the

sample to be loaded as a small bead on the tip of the pin, and it is deposited on a solid surface once the pin comes in contact with it. The pins vary in shape, from convex to flat to concave. One loaded sample can print only a few spots, which is why these pins are generally only used for low-density arrays. They do offer a few advantages, however. If sample solutions are very viscous, they can clog the split-pin types, which is not an issue for solid pins. Cleaning of these pins is also much easier, so arrays are typically very reliable between batches.

The most common type of pin that is used today is the split pin. These pins contain small (60-100 μm) slits in the center of the pin that allow a greater volume of sample to be loaded in each dip with capillary action, greatly increasing throughput. However, because the same pins are often used to handle multiple samples, they have to be carefully washed between samples. This typically includes a sonication bath step between samples to assure that all sample liquid is removed. Some disadvantages need to be considered, however. Although made of metal (typically solid steel, tungsten, or titanium), the very fine tips can still be subject to deformation from the tapping force and lead to uneven spotting. Dust and contaminants as well as highly viscous solutions can lead to clogging as well. This can be resolved by using a wider slit, but spot size reproducibility is affected. Choosing the right pinhead will depend on the sample, so careful consideration must be applied to achieve optimal conditions.

The actual printing system involves a pin head (holding anywhere from 16 to 96 pins) attached to a translation system, a sample plate, and a flat plane with the carefully positioned solid surface substrates (Figure 1.2). The pins are free to move vertically while in the pin head, so there is flexibility when they come in contact with the solid

surface. The translation machinery moves the pin head to the source plate, where the pins are lowered to collect the sample through capillary action. The sample is held by surface tension either on the tip (solid pins) or within the capillary (split pins). The pin head is then positioned over the solid surface and gently, but reproducibly, tapped, dispensing the sample onto the solid surface where it is immobilized. Spot sizes are defined by many factors, including sample viscosity, pin velocity, substrate surface properties, precision of the robotic controls, humidity, and temperature. Although pin printing represents the most popular form of microarray fabrication, it can be time consuming. For example, a single 384-well source plate printing can take up to 6.4 hours, so other spotting methods were developed to overcome this limitation.

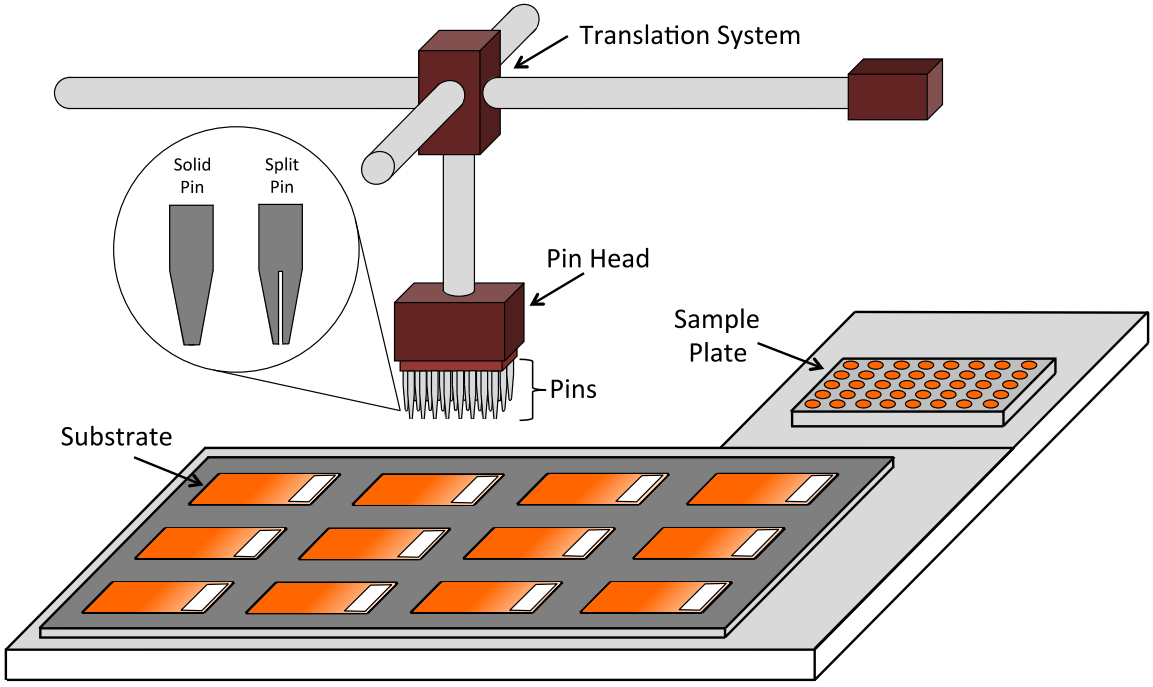

Figure 1.2. General diagram of a contact printing apparatus. A pin head, which typically holds between 16 and 96 pins, is attached to a translation system, which can accurately and reproducibly move in 3-D space. The pins are first dipped into a sample plate, where the sample adheres to the pins with capillary action. The pins are then moved above the arranged substrates and spotted by physically tapping the surface. As a serial printing technique, this is generally time-consuming, as a single 384-well source plate can take over 6 hours to fabricate.

Microstamping is a serial contact printing technique that is both simple and cheap enough to be performed in any laboratory, while allowing for the high-throughput fabrication of protein microarrays. The basic protocol involves a protein sample being adsorbed onto a patterned surface and then transferred to a substrate using physical contact. Obtaining good contact is critical in the sample spotting, so elastomeric materials, such as poly(dimethylsiloxane) (PDMS) are generally used because they conform to solid surfaces under an applied load. These elastomeric stamps are produced using a micromolding technique that requires a single photolithography step. A master mold is created by photolithography, followed by casting of an uncured elastomer onto the mold. After curing, the stamp is removed from the mold, where it can be used to adsorb to samples. Through this process, many disposable stamps can be easily fabricated, eliminating the need for the washing between samples that is required for pin printing. The major disadvantages of the stamping method involve the amount of sample that can be transferred. The sample volume transferred from stamp to substrate is not well controlled and depends greatly on the surface and sample properties. Moreover, initial sample volumes are much larger than in pin printing because only a small fraction adsorbs to the stamp and is deposited on the substrate, which can be a serious limitation when dealing with precious samples (such as patient tissues). Spots are also much larger than what can be produced using pins (300 μm vs. 60 μm), which leads to decreased spot density. Likewise, the washing of microstamps for reuse is more difficult and tedious than for pins due to the non-specific adsorption that can occur.

Inking is the first step performed in microstamping and begins by treating the hydrophobic PDMS stamp with a hydrophilic film to insure sample adherence. Similar to

a standard ink and rubber stamp, the microstamp is then simply inked by dipping into a sample well, where hydrophilic interactions mediate adsorption onto the stamp. In order to obtain uniformity in the amount of sample that is adsorbed to the stamps sites, some new techniques have been developed. Injecting devices called piezoneedles have been used to deliver equal amounts of DNA solution to stamp sites. This technique exhibits good control of spot size and is able to produce multiple arrays from a single inking step. Another technique involves loading the sample solution into wells on the back of a stamp. The solutions flow through microchannels down to the stamp tips through surface tension, and are deposited onto the substrate when careful pressure is applied.

Two types of stamping are used, direct and indirect. Direct stamping involves the stamp directly interacting with a biological sample before being brought into physical contact with the substrate, typically for a few seconds to ensure good conformation. The stamp is then removed, leaving the sample behind. Using the necessary speed, contact angle, surface roughness and spacing between stamp sites is needed to achieve uniform shape and volume in spots. One type of direct stamping technique involves a microfluidic design, which contains both inlet and outlet ports for which the biological sample can be channeled. A stamp is pressed onto the substrate and the sample solution flows through the channel and prints a spot onto the substrate. This process can be used to print incredibly concentrated samples because the number of desired molecules can be immobilized to the substrate without washing away unwanted materials.

In indirect stamping, the stamp is inked with chemical groups or self-assembled monolayers (SAMs) rather than with biological samples. Typically SAMs are patterned on a gold substrate with alkanethiols, in which a head group adsorbs onto the gold

surface, followed by a long polymer carbon chain, and ending with a reactive tail group. The stamp first prints an array using a particular alkanethiol (AT1) onto a gold substrate using traditional contact methods. Next, the remaining un-modified gold surface is coated with a secondary alkanethiol (AT2). Samples are designed so that they selectively react only with the AT1 head group and an array is formed. This type of printing is not suitable if more than one sample is to be arrayed, as non-specific adsorption and cross-reactivity can easily occur.

Nano-tip printing is the third method of contact printing techniques that is currently employed for microarray production. These technologies are able to print spots at the sub-micron level through the use of atomic force microscopy (AFM). AFM nano-tips are used in two ways: (1) adding a sample or sample binding molecules to a substrate (dip-pen lithography); or (2) removing SAM molecules from a coated surface (ATM grafting).

Dip-pen lithography uses a similar schematic to indirect sampling, but can do so at a much smaller, sub-micron level (compared to spots 300 μm in diameter obtained through microstamp printing). A sample that binds proteins is transferred to the substrate with the nano-tip in a patterned array. The remaining areas are blocked with molecules resistant to protein adsorption, as before. The arrays are finally treated with a biological sample, which only binds to the compatible surface. This has been shown to create monolayers only 30 nm wide.

AFM grafting, conversely, uses the nano-tip to remove one SAM layer (consisting of molecules resistant to protein adsorption) at selected sites, followed by treatment with reactive SAM molecules to fill in the gaps. This is useful when nano-patterns need to be

modified *in situ*, eliminating the need to repeat the entire process. The utility of this fabrication method is more limited than the other two for a few reasons. One, it is a serial technique, so it is much slower than microstamping. The slow speed reduces fabrication efficiency as well, as sample volumes must be very small because they tend to dry up during the printing process. Also, as the individual spots approach protein-size levels, non-specific binding can become a large issue. Finally, this approach, like indirect stamping, can generally only accommodate very few biological samples at a time, compared to the thousands of unique proteins that can be printed using either direct stamping or pin printing techniques. However, the greatest advantage arises from the fact that both the printing and detection of the microarrays can be performed on the same instrument, as AFM has been shown to be able to detect protein-biological sample interactions (as described later in label independent detection methods).

In contrast to contact printing, which all involved direct physical interaction between printer and substrate, non-contact printing approaches obviously do not. These techniques range from photochemistry-based methods to fluid droplet dispensing methods, but all share the same advantages: reduced contamination and higher throughput. By separating the printer from the substrate at all times, cross-contamination between samples is dramatically reduced. Additionally, the need to clean the printing devices between samples is also eliminated. Moreover, non-contact printing techniques offer the potential to increase throughput over other techniques. Many can also deposit solutions in parallel, allowing entire arrays to be produced simultaneously.

Photochemistry microarray printing involves chemical treatment of the substrate, followed by UV light exposure through photomasks. Two types of photochemistry

techniques are used, photolithography and direct photochemical patterning. In the photolithography approach, a photoresist layer is coated onto a substrate surface, followed by UV light exposure through a patterned photomask (Figure 1.3 A). This forms micrometer-sized open regions where adhesion-promoting molecules can then bind. The remaining photoresist layer is removed and the substrate is incubated with adhesion-resisting molecules.

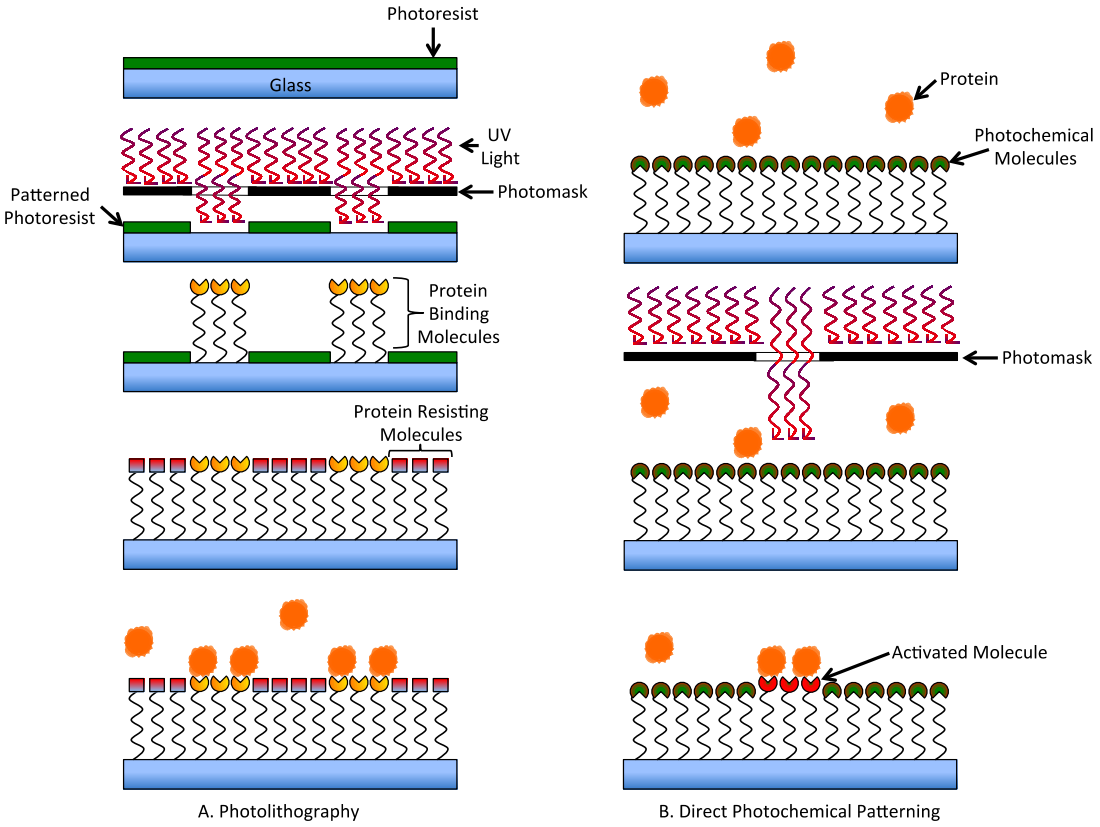

Figure 1.3. Photochemical microarray printing as a non-contact printing approach. (A) Photolithography printing first involves coating a glass surface with a photoresist coating. Illuminating the surface with UV light through a patterned photomask creates a patterned photoresist. The ablated spots are then treated with protein-binding molecules. After the remaining photoresist is removed, the remaining glass surface is coated in protein resisting molecules. Finally the protein mixture is added, where it can adhere to the array in a patterned format. (B) A similar approach, direct photochemical patterning, uses glass slides coated with photochemical molecules first in the presence of biological samples. UV light is again shone through a patterned photoresist, specifically activating these molecules, and allowing the proteins to bind. Both of these methods are limited in the diversity of samples that can be spotted, but benefit from the speed in which they can be fabricated.

Direct photochemical patterning is similar to photo-lithographic techniques, but it does not require a photoresist layer (Figure 1.3 B). Substrates are coated with photochemical molecules and activated by UV light through photomasks, where they can subsequently bind to biological molecules. These techniques are high throughput, but again, are limited in the variety of samples that can be patterned at once.

The other type of non-contact printing technique involves dispensing droplets of protein sample onto a substrate. The three most commonly used approaches are inkjet printing, electrospray deposition (ESD), and laser writing.

In an effort to reduce the cost of microarray printing machinery, attempts have been made to find cheaper alternatives in commonly available products. Oftentimes, commercially available printers are modified to dispense biological materials instead of ink. There are two commonly used inkjet printing techniques, thermal and piezoelectric. Thermal printers utilize resistive heaters (with temperatures reaching upwards of 200° C) for sample droplet dispensing, while piezoelectric printers use piezoelectric actuators to dispense droplets onto the substrate surface. Although an attractive technique, there are several disadvantages. For starters, commercially available printers are not designed to print on glass slides, so they are limited to flexible membranes such as cellulose, nylon, and nitrocellulose. These surfaces can lead to smearing, and therefore contamination between samples. Secondly, the nozzles used in inkjet printing tend to produce "satellite" spots surrounding the spot of interest, reducing printer resolution. Cleaning of the nozzles is also difficult, especially in piezoelectric printers. Finally, samples can experience high shear rates and high temperatures, regardless of the method used, which can result in the denaturing of biological samples. Some studies have shown an ability to

overcome some of these problems, namely the cleansing issue, by using deposition heads with large numbers of top loaded reservoirs (Gutmann et al., 2004). Many droplets from different samples can therefore be printed simultaneously.

Similar to inkjet printing, electrospray deposition (ESD) is a technique that has been modified from an existing application to be used with biological samples. ESD has been typically used to deposit thin films of polymers, semiconductive ceramics, and radioactive sources, but many have begun using it to deposit biological samples to a substrate for microarray fabrication. The general technique uses a dielectric mask placed between the capillaries containing biological samples and the substrate. An electrostatic field in produced between the capillary and the substrate, driving the solution out of the capillary nozzle. As the inverse charge is applied below the surface of the substrate, the biological samples are attracted to the surface, but can only pass through the holes of the dielectric mask, where they are deposited. Each capillary is filled with a different substrate, and sequentially activated after movements of the mask to different spots on the array. This allows for fast and parallel fabrication of microarrays, as well as the production of very small spots (2-6 μm). However, although spot size can be small, the distance between spots (called "pitch") is relatively large, on the order of 1 mm. This limits the overall number of spots that can be arrayed onto a single substrate slide. Another limitation arises from differences in droplet distribution from the spraying technique. Generally, only those spots directly below the capillary produce high-density spots, but surrounding spots are generally quite irregular in regards to density. However, because the pitch is so large, cross-contamination is generally not an issue. High shear rates can be damaging as well, as the solution becomes charged due to the electrostatic

field. This field can cause deformations in proteins, as well as alterations in solution pH, so it is important to take into consideration when selecting samples to use with this method.

The final method of non-contact printing uses laser ablation to produce protein microarrays, either directly or indirectly. In direct writing, a sample is mixed with glycerol and buffer and coated on a quartz disc. As a pulsed laser is scanned across the disc and local evaporation occurs, producing microscopic droplets that fall to the substrate surface. The droplets produced are smaller than one can get using traditional pin methods (50 μm) and does so with very little sample. Using this technique, over 16,000 arrays can be produced with only 500 nL of starting material. Indirect writing is a process that is similar to many of the other "lift-off" techniques mentioned before. A photoresist layer is placed on top of a substrate and is selectively removed with a photomask and laser, leaving patterned pockets with which to apply biological samples.

While all these non-contact printing methods hold great promise for the production of sub-micron spots, all require high shear, high temperatures, or both. Therefore careful consideration must be had when choosing to use these approaches.

In summary, there have been many attempts by various labs to create a fabrication system that optimizes uniformity, minimizes sample volume, and does so without contamination of the biological samples. Pin printing remains the most popular player in this process, as it overall provides the most reproducible arrays with little maintenance, albeit with the highest cost. Many of the ablation techniques can produce very small spots, but are severely limited in the number of unique samples that can be spotted on one array. Non-contact techniques eliminate many of the contamination concerns present in

pin and stamp printing, but at the cost of high temperatures and shear forces. As these technologies inevitably advance, these shortcomings have promise to be overcome.

1.4 Detection Methods

The detection methods used for protein microarrays are another important design parameter. There are two classes of detection: Label dependent and label free.

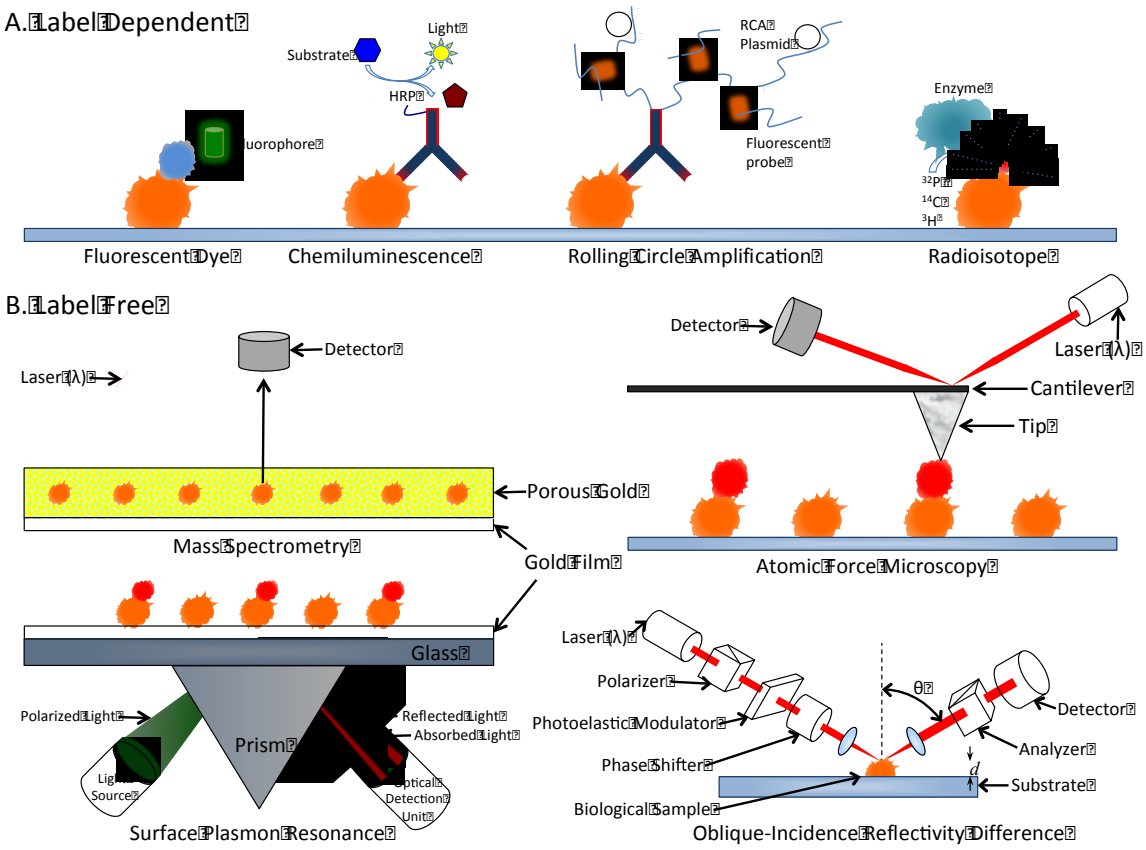

Figure 1.4. Detection methods used in protein microarray experiments. (A) Label dependent detection can alter protein activity, but provide great sensitivity. Fluorescent dyes are among the most commonly used, and can utilize multiple colors for detection of multiple interactions (i.e., acetylation and protein levels, concurrently). Chemiluminescent techniques can also be used, such as horseradish peroxidase (HRP). RCA approaches are useful with low abundance proteins, as they can greatly amplify the signal. Radioisotopes are useful for *de novo* assays and are preferred when specific antibodies are not available. (B) Label free methods eliminate the possibility of altering protein activity, but at the cost of expensive equipment and lower throughput. Mass spectrometry approaches can be used with complex mixtures. Atomic Force Microscopy (AFM) can detect ligand binding by sensing a change in height as a cantilever moves across the array. Surface Plasmon Resonance (SPR) is used to detect binding in real-time by observing changes in incidence light from the resonance of the surface plasmon when a ligand binds to and dissociates from its target. Oblique-Incidence Reflectivity Difference (O-IRD) detects changes in reflectivity between p- and s-polarizations, and can likewise be used to detect changes in biological samples from ligand binding in real-time.

Label Dependent

Several types of label dependent detection methods have been developed and optimized (Figure 1.4 A). The first is the use of fluorescent dyes with narrow excitation and emission spectra, such as Cy3 or Cy5. These are most commonly used for their convenience and wide detection range. Furthermore, they can also provide a multi-color system for multiplex assay design (Fu, 2007). Enzymatic methods can also be used to enhance signal amplification. The most common is horseradish peroxidase, but others include a modified rolling circle amplification (RCA) label that uses DNA primers attached to antibodies to create binding partners for fluorescent, complementary oligonucleotide probes. This is especially useful for the detection of low abundance proteins, where the sensitivity of chemiluminescence and traditional fluorescence is too low (Schweitzer et al., 2002). Other assays that involve enzymatic reactions on the immobilized proteins can use radioisotopes (e.g., ^{32}P, ^{33}P, and ^{14}C) for *de novo* detection. While this can provide superior signal-to-noise ratios, and can be the only reliable detection method for modifications without high-affinity antibodies, many oppose this method due to safety concerns (Chen et al., 2006).

Label Free Detection

Although useful, labeling processes can alter protein activity, so many label free methods have been developed (Figure 1.4 B). Mass spectrometry has been used for the detection of ligand binding, using MALDI-MS, SELDI-TOF-MS, and MALDI-TOF-MS

approaches (Gavin et al., 2005; Diamond et al., 2003; Evans-Nguyen et al., 2008). This approach can be done quickly and simply and with very little sample, and can also be used to directly detect analytes bound from complex samples such as urine, serum, plasma, and cell lysates (Zhang, 2012). Atomic force microscopy (AFM) is another approach that can identify analytes bound to an array by detecting a change in height of the samples on the array, which leads to measurable binding interactions (Yan et al., 2003). Another advantage of label-free methods arises from the ability to detect real-time dynamics of protein interactions. Surface plasmon resonance (SPR) has been used to analyze bimolecular interactions in real-time and has been adapted for the protein microarray (Wegner et al., 2003; Unfricht et al., 2005). Incident light can resonate with plasma on a metal surface in total internal reflection, causing resonance signals to change when ligands bind to (and dissociate from) proteins on the array surface. In this way, binding events can be monitored in real-time and kinetic parameters can be calculated. As an extension of traditional SPR techniques, SPR imaging (SPRi) combines the sensitivity of SPR with the spatial capabilities of imaging. A CCD camera is used to capture the light reflected from the entire surface and can be quantified by subtracting the background signal from the image taken before the experiment begins. As samples flow over the surface of the array, multiple images are acquired, and reflectivity changes can be quantified in real-time with a visual output as well (Kodoyianni, 2011). Similarly, the oblique-incidence reflectivity difference (O-IRD) approach allows for the very sensitive detection in the changes in reflectivity between the p- and s-polarizations. It is also used for acquiring kinetic parameters by detecting the tiny changes in the physical properties of a biological sample, such as thickness and density (Landry et al., 2008).

1.5 Basic Research Applications

Among the growing fields of scientific research that utilizes the protein microarray, basic research has been at the forefront of this technology. There is an incredible amount of diversity in the applications of this technology, which we will highlight in this section, and while these studies all involve the same general schematic, they all arrive at vastly different conclusions (Figure 1.5).

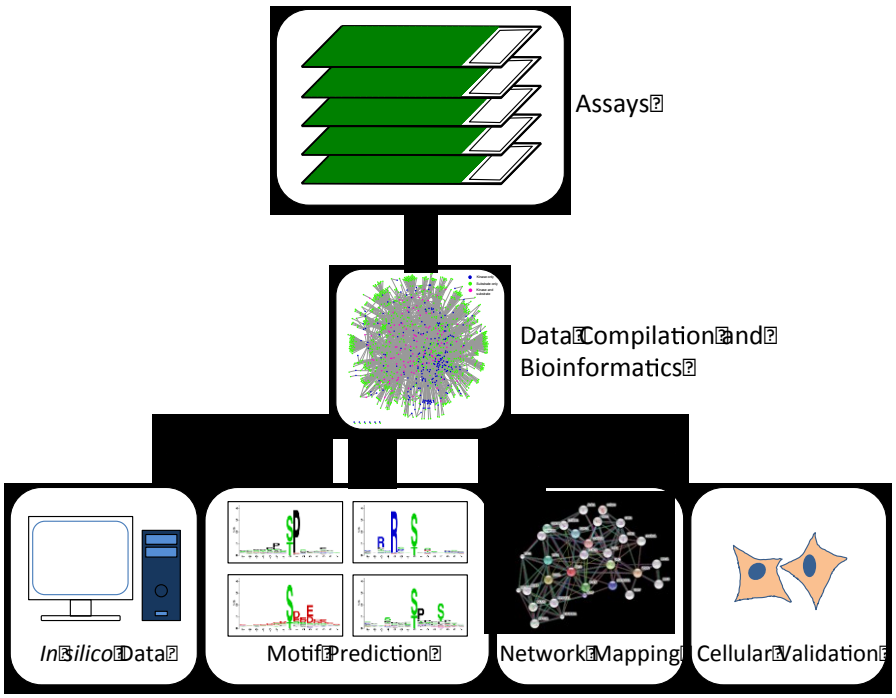

Figure 1.5. General schematic for microarray studies. A typical microarray study involves three steps; (1) Completion and detection of assays; (2) Compiling data for bioinformaticians, where biologically relevant connections are first made; (3) validation and further observations. *In silico* conclusions can be drawn through statistical algorithms, but all require further validation. Many motifs types are commonly predicted from the commonalities between hits, such as binding motifs and PTM motifs. Interactions can be combined with known data to formulate dense networks. The final step generally requires some sort of *in cellulo* or *in vivo* validations. One common weakness of all microarray studies arises from false-positives and false-negatives that can arise from any *in vivo* system, so a combination of bioinformatics and cellular validation studies helps eliminate these biases.

These studies include the detection of protein-binding interactions (protein-protein, protein-lipid, protein-peptide, protein-DNA, protein-RNA, protein-small

molecule, and protein-glycan), detection of prost-translational modifications (phosphorylation, ubiquitylation, acetylation, nitrosylation, sumoylation, and glycosylation), and profiling of monoclonal antibody specificity.

Detection of Protein-Binding Interactions

One of the most basic fundamentals of cell signaling involves the physical interaction between biological molecules. Identifying these interfaces is critical in developing our understanding of how our increasingly complex world works. Many early studies have used two-hybrid screens to elucidate protein-protein and protein-DNA interactions (Vidal et al., 1996; Young, 1998; Joung et al., 2000). This technique is a popular approach due to its low-tech nature and scalability, therefore is often used as a first step in identifying novel interactions. However, there is criticism due to the many false positives and false negatives that can arise for a variety of issues. As fusion proteins, some interactions may be inhibited due to the non-native nature of the protein. Hybridization also takes place in the nucleus, limiting the breadth of the studies to those proteins that localize to the nucleus. Protein complex purification coupled with mass spectrometry analysis is another approach that has been established to identify novel protein-protein interactions (Krogan et al., 2006). While this method has proved to have high fidelity, it is both time-consuming and not suitable for low abundance proteins. These issues, among others, contributed to the development of protein microarray technology for the study of protein-binding interactions. These arrays can be screened in parallel quickly, reproducibly, and in high throughput. One drawback lies in the inherent

in vitro nature of the assay itself, as interactions identified on the array may not occur in the context of a cell. Therefore, hits must be further characterized in order to be truly validated. Although not without some disadvantages, protein microarrays have become one of the most widely used technologies for the discovery of protein-binding interactions.

Protein-protein Interactions

MacBeath and Schreiber performed the seminal interaction study using this technology in 2000. By utilizing the same robotic equipment used to print DNA microarrays, they were able to create high-density protein microarrays and identify protein-protein interactions, kinase substrates, and small molecule targets (MacBeath and Schreiber, 2000). As a proof of principle, they first sought to prove that this technique allowed for the functional properties of the immobilized proteins to remain intact. Three pairs of known interactors were used to test this hypothesis: Protein G and immunoglobulin G (IgG); p50 (of the NF-κB complex) and its inhibitor I-κBα; and the FKBP12-rapamycin binding (FRB) domain of FKBP-rapamycin-associated protein (FRAP) and the human immunophilin FKBP12 (12kD FK506-binding protein). The first protein of each pair was immobilized on a glass surface followed by incubation with a fluorescently labeled pair. Each pair utilized a different fluorophore and they were able to show that they each bound only to their respective interactors, even simultaneously. Another test involved printing 10,800 spots onto a glass slide, with 10,799 of them containing Protein G and one specifically printed with FRB. Using a mixture of both

blue (IgG) and red (FKBP12) probes, they were able to show only one spot fluoresced red, only where the FRB was spotted. This initial study signified that the immobilization of the proteins onto derivatized slides did not affect their binding properties.

Shortly thereafter, Zhu et al. analyzed protein-protein interactions using test ligands labeled directly or indirectly with fluorescent dyes (Zhu et al., 2001). Using the first "proteome" microarray mentioned before containing approximately 5,800 recombinant yeast proteins (>85% of the yeast proteome), they were able to identify the binding partners of calmodulin (CaM). After first incubating the microarray with biotinylated bovine CaM, the proteins were detected with Cy3-labeled streptavidin. Apart from identifying six known CaM targets, 33 new CaM binding partners were discovered. Further bioinformatic analysis revealed a shared consensus motif between 14 of the 39 binding partners: (I/L)QXXK(K/X)GB, where X is any residue and B is a basic residue. Not only can targets be identified easily through this approach, motifs can also be deduced, further expanding the usefulness of these arrays.

In a related study, Popescu et al. developed a protein array consisting of 1,133 *Arabidopsis thaliana* proteins to globally identify binding partners of CaM or calmodulin-like proteins (CML) (Popescu et al., 2007). Instead of using biotinylated CaM, they used CaM amino-conjugated to Alex Fluor 647 to identify 173 different targets that bound to the three CaMs and four CMLs. Like the yeast study, a large number of previously known CaM targets were identified, as well as a diverse group of transcription factors (TFs), receptor and intracellular protein kinases, RNA-binding proteins, F-box proteins, and proteins of unknown function.

Protein-lipid Interactions

Apart from identifying CaM binding partners, Zhu and colleagues also identified the binding partners for 6 liposomes containing different phosphatidylinositides (PIs) on the yeast proteome array (Zhu et al., 2001). The liposomes contained 1% N-(biotinoyl)-1,2-dihexadecanoyl-*sn*-glycero-3-phospho-ethanolamine, triethylammonium salt (biotin-DHPE) and were similarly detected using a streptavidin-labeled Cy3 fluorescent dye. The six liposomes were able to identify 150 different protein targets, including integral membrane proteins, anchored proteins, prenylated proteins, and lipid metabolism enzymes. The binding interactions could also be characterized as "strong-binding" or "weak-binding" by comparing the fluorescent signal obtained by the liposomes to the fluorescent signal obtained from GST probing (this standard technique uses a Cy5 labeled antibody targeting the tag of the immobilized proteins on the array, yielding the relative protein concentration at each spot). Through this analysis, they were able to correlate strong binders with membrane-associated proteins and that strong phosphoinositide-binding proteins preferentially bound a specific PI. Not only could binding interactions be identified, but conclusions can be drawn concerning the binding affinities of the various PIs. This technique of comparing the binding/modification levels to the protein levels is one that is utilized often, and is another reason why protein microarrays are one of the more versatile platforms available.

More recently, Lu et al developed another fluorescent liposomal assay to identify binding partners for phosphatidylinositol 3,5-bisphosphate ($PI(3,5)P2_2$) using the same yeast proteome array used by Zhu (Lu et al., 2012). A special non-quenched fluorescent

(NQF) liposome was developed that eliminates the need to lyse the liposomes in order to see fluorescent signal. Using these modified liposomes containing ($PI(3,5)P2_2$), 162 binding proteins were identified, 22 of which were novel and shared similar functional roles with known interacting partners, including vesicle-mediated transport machinery, GTPases, and protein kinases. Collaboration with a bioinformatician revealed a novel motif, HRDIKP(E/S)N(I/L)LL, that was significant among the 162 novel ($PI(3,5)P2_2$) binding partners. While the hypothesis was that this was not the site of PI binding, it represented a novel motif among PI-interacting proteins. This approach exemplifies a beneficial improvement over traditional fluorescent liposomal nanovesicles, as it can be used on microarrays, and has promising applications in future high-throughput studies of protein-lipid interactions. It is also important to note that with large studies such as this, it is crucial to have strong relationships with bioinformaticians to aid in the analysis of the large amounts of data that is produced.

Protein-peptide Interactions

Jones and colleagues developed a unique array that used protein domains instead of full-length proteins in order to investigate protein-peptide interactions deemed important in the ErbB family of receptors (Jones et al., 2006). The array was printed with 106 Src homology 2 (SH2) and 41 phosphotyrosine binding (PTB) domains, along with several tandem domains. Arrays were incubated with 61 peptides representing tyrosine phosphorylation sites on the four ErbB receptors. In order to quantitatively measure the protein-peptide interactions, eight concentrations of each peptide (between 10 nM and 5

mM) were tested, leading to the following insights: (1) the study of these interactions in a non-competitive format revealed high-affinity binding sites for both SH2 and PTB domain, but do not imitate the consensus recognition sequences; (2) ErbB2 recruitment sites are more promiscuous than any of the other receptors; (3) when comparing the highest-affinity interactions, the proteins that bind to EGFR comprise a small subset of those that bind to ErbB3; and (4) at high concentrations of EGFR and ErbB2, the binding becomes much more promiscuous, which may contribute to the high oncogenic potential that has been seen in various cancers.

Protein-DNA Interactions

Another common application of protein microarrays is in the characterization of protein-DNA interactions (PDIs). One of the earliest studies screened novel DNA-binding proteins by probing the yeast proteome microarray with fluorescently labeled yeast genomic DNA (Hall et al., 2004). Greater than 200 proteins were identified, and half of those were not previously known to bind DNA. In order to determine whether the novel DNA-binding proteins were not artifacts (i.e. they nonspecifically bound *in vitro*, but not *in vivo*), the group tested eight novel proteins by using ChIP/chip. The proteins of interest were tagged with 13 c-Myc epitopes at the C-terminal before treatment with formaldehyde to cross-link the protein to the potentially bound chromosomal DNA. The Myc-tagged protein was then immunoprecipitated from the cell lysates, where the DNA was labeled with Cy5. Five of the eight proteins showed no enrichment in any particular DNA loci, but three proteins (Mtw1, Dig2, and Arg5,6) were found to be specifically

associated with DNA loci *in vivo*. The most important of which was Arg5,6, which is a single yeast gene that encodes two enzymes involved in arginine biosynthesis. And while Mtw1 and Dig2 bound to multiple loci, Arg5,6 only bound to one specific loci found in both nuclear and mitochondrial chromatin. Deletion of Arg5,6 altered the transcript levels of both nuclear and mitochondrial target genes, signifying that metabolic enzymes can directly regulate eukaryotic gene expression.

Though shared genomic DNA can be used as probes, individually synthesized short DNA motifs can also be used as probes to interrogate the entire yeast TF repertoire immobilized on a glass slide. Indeed, the Snyder and Johnston groups were the first to establish this approach as reported in a PNAS paper in 2006 (Ho et al., 2006). Using a microarray containing 282 known yeast TFs, they were able to identify interactions with 75 evolutionarily conserved DNA motifs. Again, over 200 PDIs were identified, with greater than 60% of them previously unknown. It was shown that 15 of the proteins bound almost all of the DNA probes non-specifically, while 62 proteins bound to at least one probe. Further analysis identified Yjl103p as a novel DNA-binding protein and the DNA binding motif was found to be $CGGN_{8/9}CGG$. In order to obtain the target genes for this transcription factor, Yjl103p was either overexpressed or deleted before gene expression profiling. Over 500 genes were differentially expressed with overexpression, many of which were involved with carbohydrate metabolism, stress response, or oxidative phosphorylation. This was one of the first transcriptome-wide studies performed on an entire organism and paved the way for the analysis of the human transcriptome.

Following up on the yeast transcription factor study from Ho et al., Hu and colleagues undertook a large-scale analysis of human PDIs using a protein microarray composed of 4,191 unique human proteins in full-length (henceforth known as the TF array), including approximately 90% of the annotated TFs and a wide range of other protein categories, such as RNA-binding proteins, nucleotide-binding proteins, transcription co-regulators, mitochondrial proteins, and protein kinases (Hu et al., 2009). The protein microarrays were probed with 400 predicted and 60 known DNA motifs and a total of 17,718 PDIs were identified. Many known PDIs and a large number of new PDIs for both well characterized and predicted TFs were recovered, and new consensus sites for over 200 TFs were determined, which doubled the number of previously reported consensus sites for human TFs. Unexpectedly, over 300 proteins that were previously unknown to specifically interact with DNA showed sequence-specific PDIs, suggesting that many human proteins may bind specific DNA sequences as a moonlighting function. To further investigate whether the DNA-binding activities of these unconventional DNA-binding proteins (uDBPs) were physiologically relevant, in-depth analysis was carried out on a well-studied protein kinase, Erk2, to determine the potential mechanism behind its DNA-binding activity. Many *in vitro* and *in vivo* assays were performed, including electrophoretic mobility shift assays (EMSAs), luciferase assays, mutagenesis studies, and chromatin immunoprecipitation (ChIP). These studies showed that the DNA-binding activity of Erk2 was independent of its protein kinase activity and it acts as a transcription repressor of transcripts induced by interferon gamma signaling. Other than Erk2, many other uDBPs showed sequence-specific DNA-binding activity, and many of their consensus sequences are highly similar to those recognized by

annotated TFs. This suggests that moonlighting functions of uDBPs based on their sequence-specific DNA-binding activity may be a widespread phenomenon in humans.

Using the same approach, new proteins that may carry enzymatic activity in DNA repair can be discovered. A bacterial proteome microarray composed of 4,256 proteins encoded by the *E. coli* K12 strain (≈99% coverage of the proteome) was developed by Chen and colleagues (Chen et al., 2008). In order to identify proteins involved in DNA damage recognition, end-labeled, double stranded (ds) DNA probes containing abasic or mismatched base pairs were used. Two proteins of unknown function, YbaZ and YbcN, were identified as binding with high affinity to abasic sites and mismatched sites, respectively. Both of these proteins were shown to be base-flipping through the use of a 2-aminopyrine (Ap)-modified DNA approach. The fluorescence of 2-Ap is quenched when properly paired within the duplex, but the fluorescent intensity increases dramatically when flipped into an extrahelical conformation. An increase of more than 1,000-fold was observed for both proteins. By probing the same proteome array with purified YbaZ, it was also found that YbaZ tightly interacts with the type IV helicase HelD.

Protein-RNA Interaction

It was long thought that RNA was only used for translation (mRNA, rRNA, and tRNA), but with the discovery of over 20 types of RNA, many have been characterized and shown to have a wide variety of functions in biology. Long non-coding RNAs (lncRNAs) are most commonly thought as precursors to short RNA fragments, such as

microRNAs. However, recent studies have shown that some of these lncRNAs can be several kilobases in length, and can be conserved extensively at the nucleotide level (Guttman et al., 2009). Like microRNA, some lncRNAs block gene expression by antisense base pairing, but many act in *cis* by regulating mRNA transcription. This is achieved by modifying heterochromatin formation near genomic loci. Nonetheless, there have been relatively few lncRNAs functionally characterized, especially those which appear to act in *trans* to regulate gene transcription. While there are existing methods available to study protein-RNA interactions, they are laborious and require large amounts of cell material. Protein microarrays provide a unique and high-throughput platform from which these interactions can be quickly identified.

Rapicavoli and colleagues were among the first to do so when they sought to further characterize the long non-coding RNA *Six3OS*, which is co-expressed with the homeodomain factor Six3 (Rapicavoli et al., 2011). Six3 plays a pivotal role in mammalian eye development, where it regulates both early eye formation and cell specification to the postnatal retina. Utilizing a human protein microarray, they were able to demonstrate that both mouse and human *Six3OS* was able to bind to five proteins, including Eya1, a homologue of the *eyes absent* gene of *Drosophila*, and Ezh2, a chromatin remodeling enzyme. It was proposed that *Six3OS* modulates the expression of Six3 target genes by acting as a transcriptional scaffold that recruits histone modifying enzyme complexes, acting in a *trans* manner to do so.

More recently, a microarray containing more than 9,400 human proteins was used as a platform to identify binding interactions between proteins and 10 full-length coding and noncoding RNAs (Siprashvili et al., 2012). The 137 protein-RNA interactions

discovered were enriched for known human RNA binding domains, including RRM, RBD, K homology domains, and zinc finger motifs. One such protein, Staufen 1 (Stau1), was identified as a binding target of *TP53* mRNA. Using *in vivo* RNA-pulldown assays, they were able to show that Stau1 bound sense-*TP53* mRNA, but not control mRNA. This was validated using reciprocal pull-down experiments. Further characterization showed that Stau1 binding to *TP53* RNA aided in maintaining its RNA levels during transcriptional blockades.

Protein-small Molecule Interactions

Protein microarrays have also been shown to be useful in discovering new drug molecules and drug targets. Rapamycin is a small molecule drug that can induce a starvation response and inhibit cell growth through its target TOR (target of rapamycin), a highly conserved protein kinase regulating cell proliferation and metabolism, both in yeast and in humans. Huang et al. identified small-molecule inhibitors of rapamycin (SMIRs) and small molecule enhancers of rapamycin (SMERs) (Huang et al., 2004). They obtained the binding partners of two SMIRs, SMIR3 and SMIR4 by probing the entire yeast proteome with biotinylated versions of the SMIRs, and two targets for rapamycin inhibition were identified. The first, Tep1p, is a homolog of the mammalian PTEN tumor suppressor; the second is Ybr077cp (Nir1p), a protein previously with no known function. Both Nir1p and Tep1p were shown to associate with PI(3,4)P2, which suggests a novel mechanism by which phosphatidylinositides might modulate targets of

the rapamycin pathway and the ability of protein microarrays to identify specific protein-small molecule interactions.

Polyanions are a particular class of proteins with considerable negative charge and are hypothesized to be involved in intracellular organization, protein stabilization and folding, and protein transport. Salamat-Miller et al. used the yeast proteome array to probe for binding partners of well-characterized polyanionic proteins: actin, tubulin, heparin, heparin-sulfate, and DNA (Salamat-Miller et al., 2006). By using biotinylated versions of these five polyanions, 893 polyanion-binding proteins (PABPs) were identified. The polyanions and their binding partners were shown to form a network involved with maintaining the structure and activity of yeast cells.

Protein-glycan Interactions

One of the integral post-translational modifications involved in cellular membrane formation is glycosylation. Proper glycosylation is critical in dictating proper conformation of many membrane proteins, retaining stability of some secreted glycoproteins, and facilitating cell-cell adhesion. The yeast proteome array was used to further examine the roles of protein glycosylation in yeast (Kung et al., 2009). Since the proteins were all purified in their native host, they were expected to maintain most of their PTMs. To that end, the arrays were probed with two fluorescently labeled lectins: Concanavalin A (ConA), which recognizes mannose, and Wheat-Germ Agglutinin (WGA), which recognizes N-acetylglucosamine (GlcNAc). Two separate yeast proteome arrays were used in this experiment. The first was the aforementioned array used in

many of the previously discussed studies developed by Zhu et al. that used a GST affinity tag on the N-terminal and contained ≈5,800 yeast proteins. The second array contained ≈5,600 yeast proteins with a C-terminal Protein-A IgG-binding domain and was developed by Gelperin and colleagues (Gelperin et al., 2005). There is an important distinction between these two arrays, especially as it pertains to glycoprotein signaling. Proteins that are anchored in the membrane at the C-terminal (and some type II proteins) are optimal for the N-terminal tagged array (Zhu) because the signaling sequence and transmembrane domains are present at the C-terminal and are unlikely to be effected by an N-terminal tag (e.g. cytochrome *b5* and the SNARE proteins). In contrast, proteome chips with C-terminal protein fusions (Gelperin) are optimal for studying type-I, and most of type-II and type-III membrane proteins that are glycosylated through the secretory pathway; the signaling and modifications with N-linked glycans occur at the N-terminal and would be hindered by an affinity tag at the N-terminal. ConA and WGA identified 124 and 174 proteins, respectively, in the N-linked array and ConA and WGA identified 236 and 142 proteins, respectively, in the C-linked array. In total, 534 proteins were identified, 406 of which were previously not known to be glycosylated. Gene ontology analysis identified enrichment in proteins from the secretory pathway, TFs, and other mitochondrial proteins. These mitochondrial protein targets were further examined by treating yeast cell cultures with tunicamycin, an inhibitor of N-linked protein glycosylation. Two of the four mitochondrial proteins identified showed partial distribution to the cytosol and reduced localization to the mitochondria, suggesting a new role of protein glycosylation in mitochondrial protein function and localization.

Another approach to identifying protein-glycan interactions involves the inverse layout of the microarray. A high-content lectin microarray consisting of 94 unique commercial lectins was fabricated and used to profile accessible surface glycans of mammalian cells (Tao et al., 2008). Twenty-four human cell lines were labeled and applied to the microarray. Each cell line was subjected to a binary algorithm that was developed to generate "glycan signatures," resulting in hierarchical clusters based on each line's accessible glycan composition. By comparing the glycan profiles of a breast cancer cell line and its cancer stem-like cell derivatives, three lectins (LEL, AAL, and WGA) were found to specifically recognize MCF7 cells, but not the derivatives. To further validate this results, the authors employed LEL-conjugated beads to purify away the normal MCF7 cells from the cancer stem-like cells (estimated as ≈0.1% in the cell population) in order to enrich for cancer stem-like cells. Next, by using a mouse model to test the enrichment of the cancer stem-like cells, they were able to show that two weeks following injection of the LEL-depleted cancer stem-like cell enriched cultures, the average tumor size was greater than two-fold bigger than the control group injected with a similar number of normal MCF7 cells. This study demonstrates the utility of using lectins on a microarray to identify novel cell surface markers on cancer stem-like cells, and the ability to enrich a sample for cancer stem-like cells.

The affinities for lectins are typically very low (*Kd* is in the range of $10^{-3} – 10^{-6}$ M), which can pose some difficulties when performing a cell-binding assay because low-affinity interactions may be washed away from the immobilized lectins. This is especially true when dealing with live cells. To overcome these problems, several researchers have modified this technology to improve the strength of the binding

interactions. One example involves the antibody-assisted lectin profiling (ALP) approach developed by Kuno et al. for detecting glycoproteins at low concentrations (Kuno et al., 2009). This method was used to analyze the glycan structures of the platelet aggregating factor hPod, which has been proposed to enhance the metastatic potential of glioblastoma cells. The hPod protein complex was first enriched by immunoprecipitation, before incubation on a lectin microarray to identify its associated glycans. The additional modification of the platform involves the use of an evanescent-field activated fluorescence detection system, which allows for a label-free, real time detection system. An evanescent field is generated within 200 nm of the solid surface, rendering the background signal so low that washing steps are not necessary. Additional studies have shown that this system is by far the most sensitive detection system among lectin microarrays, reaching a reported detection limit in the 100 pM range (Uchiyama, 2006).

Finally, Li and Tao et al. reported a two-phase discovery and validation scheme improving the sensitivity of the lectin microarray in the study of prostate cancer biomarkers (Li and Tao et al., 2011). The approach first involves pooling tissue samples from four groups in equal amounts (50 μg): normal, nonaggressive cancer, aggressive cancer, and metastatic cancer. The discovery phase extracts prostate specific antigen (PSA) and membrane metallo-endopeptidase (MME) proteins from each tissue group using an anti-total PSA antibody and an anti-MME mAB, respectively. The immunoprecipitated PSA and MME proteins were incubated on a lectin microarray, followed by detection of PSA and MME proteins using anti-PSA and anti-MME mAbs, respectively. Comparison of signals between each group of pooled tissue revealed that the fraction of PSA that is O-glycosylated (as recognized by jacalin) or Neu5Ac-

conjugated (as recognized by SNA) was highly elevated in aggressive prostate cancer and metastatic prostate cancer groups. It was shown also that the fraction of MME that was modified by either GalNAx or GlcNac showed this elevation as well. Confirmation of these finding were shown using an immunosorbent assay, in which PSA and MME were first captured on an ECL plate coated with anti-PSA and anti-MME mAbs, followed by detection with biotinylated lectins. These studies show the power and adaptability of protein microarrays, even as they involve the detection of comparably weak protein-glycan or protein-lectin interactions.

Interaction Type	**Array Content**	**Type of Probe**	**Reference**
Protein-protein	60 EBV viral proteins	Human protein	Zhu et al., 2006
	4,191 human proteins	Viral protein	Shamay et al., 2012
Protein-lipid	Yeast proteome	$PI(3,5)P2_2$ liposomes	Lu et al., 2012
Protein-peptide	159 human SH2 and PTB domains	Peptides	Jones et al., 2006
Protein-DNA	282 yeast TFs	DNA motif	Ho et al., 2006
	4,191 human proteins	DNA motif	Hu et al., 2009
Protein-RNA	9,400 human proteins	Coding and noncoding RNAs	Siprashvili et al., 2012
	Yeast proteome	BMV RNA loop	Zhu et al., 2007
Protein-small molecule	Yeast proteome	Small molecule inhibitors/enhancers of rapamycin	Huang et al., 2004
Protein-glycan	Yeast proteome	Lectins	Kung et al., 2009
	94 lectins	Live mammalian cells	Tao et al., 2008

Table 3. Applications of Microarrays to Detect Protein-binding Interactions.

Detection of Post-Translational Modifications

Protein post-translational modifications (PTMs) are one of the most important mechanisms used by cells to directly regulate protein activity. Hundreds of PTMs have been identified, and are enzyme-dependent and reversible, including protein (de)phosphorylation, (de)ubiquitylation, (de)SUMOylation, (de)acetylation, (de)nitrosylation, and (de)glycosylation. Many labs have attempted to further understand the biological consequences of these PTMs, as it is important to identify downstream targets at a systems level. "Shotgun" MS/MS techniques have been a workhorse in the identification of PTMs in mammalian proteomes. However, this bottom-up approach does not identify the upstream enzymes responsible. The functional protein microarray provides a fantastic platform for the investigation of the players responsible for this complex category of cell signaling.

Protein Phosphorylation

Protein phosphorylation is one of the better-studied PTMs, and as such is considered one of the central players in most, if not all, cellular processes. The application of protein microarrays to the study of phosphorylation was first demonstrated by Zhu et al. in 2000. Seventeen different substrates were immobilized on a nanowell protein microarray, followed by individual kinase assays with 119 out of the 122 known yeast kinases (Zhu et al., 2000). The *in vitro* phosphorylation was detected by using ^{33}P-γ-ATP as the phosphate donor in the kinase assay, and signal was acquired with either

film or a phosphorimager screen, which can be used to quantify the isotopic signal. By using this approach, they were able to determine the kinase substrate specificity of the yeast kinome and identify a large number of tyrosine kinases.

As a follow up to this project, the Snyder group performed a large-scale "Phosphorylome Project" using the aforementioned yeast proteome arrays (Ptacek et al., 2005). Eighty-seven yeast kinases were purified and incubated with the yeast proteome microarrays in the presence of ^{33}P-γ-ATP. A total of 1,325 distinct protein substrates were identified, representing a total of 4,129 phosphorylation events. Some observations were made immediately regarding the global kinase activity of yeast. Most (73%) substrates were recognized by fewer than three kinases, signifying the strong specificity of kinases for their respective substrates. Transcription factors were the largest class of proteins that were phosphorylated, supporting the regulatory role of phosphorylation on protein expression. Further analysis provided a global network that connected kinases to their potential substrates, offering new opportunities to identify new signaling pathways and cross-talk between pathways.

The natural evolution of this research reached higher eukaryotes in early 2013 when Newman et al. analyzed human kinase-substrate relationships (KSRs) using the TF microarray developed in the lab previously (Hu et al., 2009) which contained 4,191 unique, full-length human proteins. To date, approximately 2,000 human KSRs have been experimentally verified; that is, a known kinase has been shown to phosphorylate a known substrate. This pales in comparison to the over 70,000 phosphorylated serine, threonine, and tyrosine residues that have been identified *in vivo* through mass spectrometry (MS/MS). However, as mentioned before, simply knowing which residues

are modified does not give any insight to the pathways or even the direct upstream kinase that is responsible for the modification. A new strategy, dubbed CEASAR (Connecting Enzymes And Substrates at Amino Acid Resolution), was employed to fabricate a high-resolution map of human phosphorylation networks, connecting kinases to their immediate downstream targets (Newman et al., 2013). The TF microarrays were incubated with 289 unique, active, and full-length human kinases in the presence of ^{32}P-γ-ATP. These experiments yielded 24,046 phosphorylation events involving all 289 kinases and 1,967 unique substrates to create a "rawKSR" dataset. Using a Bayesian statistics model, the raw KSR dataset was refined by hypothesizing that the KSRs would be more physiologically relevant if they were known to share the same tissue specificity and sub-cellular localization, and if they were already known to interact from literature. A high-confidence dataset was created containing 3,656 refined KSRs (refKSRs) involving 255 unique kinases and 742 substrate proteins. Finally, a combined (comKSR) dataset was created that included both the refKSR data and 719 known KSRs. Validation studies in HeLa cell lines showed that 76% of tested KSRs showed some sort of kinase-dependent change in the substrate (e.g. stabilization, degradation, or mobility shift). At the same time, an integrated algorithm was developed, termed M3 (Motif discovery based on Microarray and MS/MS), to systematically identify phosphorylation motifs. Combining the large MS/MS datasets with the rawKSR and known KSR dataset, over 13,000 of the 70,000 known phospho-sites identified by MS/MS were mapped to 1,644 substrates. Three hundred consensus motifs for 284 kinases were identified, representing 55% of the human kinome. Integrating the information from both the motif dataset and the *in vivo* phosphorylation site dataset into the comKSR dataset created a high-resolution

phosphorylation map connecting 230 kinases to 2,591 *in vivo* phosphorylation sites on 652 substrates, and identifying over 3,600 new KSRs. Using the newly generated phosphorylome map allowed an intermediate kinase to be identified in BCR signaling. Protein kinase A (PKA) was identified as both a substrate for Bruton's tyrosine kinase (BTK) and a kinase for ARID3A, filling in a gap that had not been identified before. The phosphorylations of both PKA and ARID3A were validated *in vivo*, and two phosphorylation sites on ARID3A that were predicted using M3 were verified. It was shown that BTK phosphorylates and activates PKA during BCR signaling, which in turn leads to the stabilization of ARID3A upon phosphorylation by PKA. Taken as a whole, this study represents a bright future of what functional protein microarrays can provide – a global understanding of cell signaling pathways through careful experimentation and bioinformatics.

Several smaller scale studies of kinase-substrate interactions have been reported as well. For instance, Popescu et al. probed an array containing 2,158 *Arabidopsis* proteins with 10 *Arabidopsis* mitogen-activated protein kinases (MPKs) (Popescu et al., 2008). They identified 570 putative MPK phosphorylation targets, which were enriched in transcription factors involved in the regulation of development, defense, and stress response. A commercially available human protein microarray comprised of approximately 3,000 individual proteins was used to identify substrates of cyclin-dependent kinase 5 (Cdk5), a serine/threonine kinase that plays an important role during central nervous system development (Schnack et al., 2008).

As shown previously in the protein-RNA interaction applications, viral proteome microarrays are a very useful tool in the study of host-pathogen interactions. In 2009,

Zhu and colleagues used an Epstein-Barr herpesvirus (EBV) protein microarray to investigate the function of an EBV-encoded protein kinase, BGLF4, via phosphorylation and binding assays (Zhu et al., 2009). They identified a total of 23 BGLF4 substrates and interactors, including EBNA1, a protein that is essential for the replication and maintenance of the episomal EBV genome during latency. The authors were able to show that BGLF4 acts as a negative regulator of EBNA1's replication function and raised the possibility that the induction of BGLF4 kinase activity may provide a novel means of eliminating EBV genomes from latency in infected cells.

Protein Ubiquitylation

Ubiquitylation is one of the most prevalent PTMs in eukaryotes and controls a variety of intracellular signaling events, but its regulatory mechanisms are largely unknown. The use of protein microarrays represents a rapid and high-throughput method of identifying players in this pathway, and Gupta and colleagues did just that in 2007 (Gupta et al., 2007). Rsp5 is a well-characterized HECT-domain E3 ligase from yeast belonging to the Nedd4 family. Using a commercial yeast proteome array for duplicate reactions, they were able to identify 40 high confidence Rsp5 substrates. Rsp5 contains 5 WW domains, which are known to bind to specific substrates by recognizing a (L/P)PXY sequence known as a PY motif. Not surprisingly, 72% of these substrates contained at least one of these sequences. Many of these were further validated *in vitro* and *in vivo*, indicating the fidelity of this assay. After identification of the substrates of Rsp5, they again turned to the microarray to identify its binding partners by treatment with a

fluorescently labeled Rsp5, which again showed enrichment of proteins containing a PY domain. From this, they were able to build both an interaction network and a substrate network.

At the same time, Lu et al. used a protein microarray approach for the identification of Rsp5 substrates as well, but were able to further characterize two of the substrates (Lu et al., 2008). Using the Zhu lab yeast proteome array, more than 90 new substrates were identified, and eight were validated as *in vivo* substrates of Rsp5. Characterization of one substrate, Rnr2 (a ribonucleotide reductase), revealed that Rsp5-dependent ubiquitylation affects subcellular localization. Heterozygous null mutants of Rnr2 correspond to hypersensitivity to DNA damage and treatment with the RNR specific inhibitor, hydroxyurea (HU). After DNA damage, another member of the RNR family, Rnr4, is redistributed within cells, but it is not known whether it is due to a PTM. They were able to show that Rnr2's localization was dependent on Rsp5, as it was localized to both the cystosol and nucleus in a *wt* strain, but only localized to the nucleus in an Rsp5 knockout strain. Both of these previous studies highlight the power that protein microarrays can hold in the analysis of ubiquitylation.

While each of the previous assays described were able to identify true *in vivo* interactions in an *in vitro* setting, one limitation is apparent, as none can mimic the precise cellular conditions used by these proteins *in vivo.* This can lead to false positives due to the promiscuity of the applied protein, or false negatives due to some of the necessary co-factors being absent from the reaction mix. The use of concentrated mammalian cell extracts in combination with protein microarrays can serve to identify PTM targets in a semi-*in vivo* setting while alleviating the challenge of analyzing a

complex mixture. Merbl and Kirschner generated from HeLa S3 cells three cell extracts from the two distinct phases surrounding anaphase: the mitotic checkpoint (treated with nocodazole, which prevents mitotic spindle formation and checkpoint arrest), the anaphase release phase (treated with the E2 ligase UbcH10), and an anaphase release phase containing an APC (anaphase promoting complex) inhibitor as well as UbcH10 (Merbl and Kirschner, 2009). The synchronized cell extracts were incubated with Invitrogen's Human ProtoArray (composed of 8,000 human proteins) and using the anti-polyubiquitin antibody FK1, identified 132 polyubiquitylated proteins. The authors expected to recover substrates of APC and were able to identify 11 of the 16 known APC targets on the array, confirming the experimental design. Validation studies performed in rabbit reticulocyte lysate confirmed the degradation/ubiquitylation of 7 novel APC substrates. This study demonstrates the efficacy of using protein microarrays in combination with cell extracts to recapitulate the global PTM signature in a specific cellular state.

Protein Acetylation

Acetylation of histone residues by histone acetyltransferases (HATs) and deacetylation by histone deacetylases (HDACs) has been shown to be a key regulator of chromatin structure and transcription. While histones appear to be the dominant substrate for this modification, it has been hypothesized that many other non-histone proteins may be modified by these enzymes. For example, the HAT Esa1, which is part of the essential nucleosome acetyltransferase of H4 (NuA4) complex, is the only vital HAT in

yeast, suggesting that it may target other non-histone proteins critical for survival. Non-histone substrates of the NuA4 complex were identified by Lin et al. by performing acetylation reactions on the yeast proteome microarrays (Lin et al., 2009). Arrays were treated with purified NuA4 complex in the presence of ^{14}C-Acetyl-CoA as the donor, and 91 proteins were found to be readily acetylated. Further validation of 20 targets revealed 13 with Esa1-dependent acetylation in cells. Phosphoenolpyruvate carboxykinase (Pck1p) was further characterized by mass spectroscopy, where acetylation was observed on Lys19 and Lys514. Mutagenesis at these sites demonstrated that K514 acetylation is critical in enhancing the activity of Pck1p, resulting in a longer life span for yeast growing under starvation.

S-Nitrosylation

S-nitrosylation is a unique PTM in that it is independent of enzyme catalysis, but is nonetheless an important one that affects a wide range of protein and cellular processes, including inflammation and protection from apoptosis. Recently, Foster et al. developed a protein microarray-based approach to identify proteins reactive to S-nitrosothiol (SNO), the donor of NO+ in S-nitrosylation (Foster et al., 2009). S-nitrosocysteine (CysNO), a highly reactive SNO, was added to a commercially available yeast microarray. Nitrosylated proteins were detected using a modified biotin switch technique (BST), which converts an S-nitrosothiol into an S-biotinylated Cys. The arrays were subsequently probed with an anti-biotin antibody, followed by a fluorescently labeled secondary antibody. The top 300 proteins with the strongest signal were further

characterized and showed that proteins with active site Cys thiols residing at N termini of alpha helices or within catalytic loops were particularly enriched. However, even among these protein families, there was significant variation in the S-nitrosylation, signifying that secondary structure or intrinsic nucleophilicity of Cys thiols was not sufficient to interpret the specificity of S-nitrosylation. Further analyses revealed that NO-donor stereochemistry and structure had significant impact on S-nitrosylation efficiency.

O-GlcNAcylation

Although many, if not most, proteins have been identified as phosphorylated on serine and threonine residues, modification by β-N-acetylglucosamine (*O*-GlcNAc) on serine and threonine residues has emerged as another fundamental regulatory mechanism in cell signaling. *O*-GlcNAc cycling on proteins is mediated by two enzymes: *O*-glycosyltransferase (OGT), which glycosylates S/T residues, and *O*-GlcNAcase (OGA), which hydrolyses S/T residues back to their native state. Like phosphorylation, *O*-GlcNAcylation can modify protein function, including kinase activity, turnover, protein-protein interactions, subcellular localization, DNA affinity, and transcription activity. Dias et al. were the first to use a functional protein microarray to probe for O-GlcNAc on a large scale (Dias et al., 2012). A commercially available kinase array was used which contained 152 full-length human kinases. Using tritiated UDP-GlcNAc (UDP-^{3}H-GlcNAc) as the donor along with purified OGT, they were able to identify 42 kinases modified *in vitro*. Further validations in HEK293 cells were performed that confirmed the modification occurs *in vivo* as well. While phosphorylation is controlled in humans

by over 500 kinases and over 150 phosphatases, there is only one OGT and one OGA, suggesting that the specificity with which these enzymes react most likely involves other binding partners, and further studies will need to be performed in order to elucidate the breadth of this modification.

PTM Studied	**Substrate**	**Enzyme**	**Reference**
Phosphorylation	Yeast proteome	Yeast kinases	Ptacek et al., 2005; Zhu et al., 2000
	3,000 human proteins	Human CDK5	Schnack et al., 2008
	Human TF array	Human kinases	Newman, et al., 2013
	Human proteome	Human CK2	Tarrant et al., 2012
Ubiquitylation	Yeast proteome	Ubiquitin E3 Rsp5	Lu et al., 2008
SUMOylation	Human	E3 RanGAP1	Oh et al., 2007
Acetylation	Yeast proteome	NuA4 complex	Lin et al., 2009
	E. coli proteome	PAT	Thao et al., 2010
S-nitrosylation	Yeast proteome	N/A	Foster et al., 2009
O-glycosylation	Human kinases	OGT	Dias et al., 2012

Table 4. Summary of Post-translational Modification (PTM) Studies Using Functional Protein Microarrays

Profiling Monoclonal Antibody Specificity

Antibodies have widespread applications in proteomic studies, but some difficulty lies in producing antibodies with sufficient specificity. Monoclonal antibodies (mAbs) are a better option than polyclonal antibodies for most applications, but tend to be much more expensive and can sometimes be too specific. Protein microarrays offer a technique to be able to test the specificity of these antibodies. After immunizing mice with live cells from human livers, Hu et al. isolated 54 hybridomas with binding activities to human cells and identified antigens for five mAbs by screening on a protein microarray consisting of 1,058 unique human liver proteins (Hu et al., 2007). The five identified mAbs were subsequently used to characterize the expression profiles of their corresponding antigens in both normal liver cells and hepatoma cells. Among them, eIF1A, an essential initiation factor in translation, was found to be present in normal hepatocytes, but not in any hepatoma cells, suggesting that liver carcinomas likely have suppressed translation.

In 2012, Jeong and colleagues used a similar approach combining immunization with live human cells and even larger microarray-based analysis to rapidly identify monospecific monoclonal antibodies (mmAbs) (Jeong et al., 2012). The protein microarray used in this study is one of the largest to date, as it contains almost 17,000 individually purified full-length human proteins. When the monoclonal antibodies were used on this array and only recognized a single antigen, they could be identified as monospecific mAbs. These antibodies were tested against their respective antigen and could be successfully identified in western blots (WB), successfully immunoprecipitate

endogenous antigen from homogenate, and successfully used in immunocytochemistry (ICC) in transfected human cell lines. The specificity with which this approach created mmAbs was more useful (e.g. more applications – WB, IP, ICC/IHC) than many other commercially available sources. For example, 66% of the mmAbs tested were IP-grade, compared to the commercial average of 18%. Future studies may be able to use this approach to tackle the issues analytical microarrays possess by easily and rapidly producing highly specific mmAbs.

1.6 Clinical Research Applications

As the basic research field continues to profit from protein microarrays, so does the clinical field. The ability to more sensitively and rapidly screen patient samples against a large array of proteins inches us closer to what many consider the future of health care – personalized medicine. While we certainly are not at the level yet, there is great promise shown in many of the studies that have arisen the last decade. In this section, we will discuss three applications of protein microarrays in the clinical field: biomarker identification, pathogen-host interactions, and cancer research.

Biomarker Identification

One of the most rapidly growing applications in the field of clinical proteomics using protein microarrays is biomarker identification. This was first used in traditional serology studies, which focused on diagnostic identification of antibodies in patient

serum samples. These antibodies are produced as a part of the immune response to an infection, a foreign protein, or, in the case of autoimmune disease, against ones own proteins. Using protein microarrays as a platform for potential antigens, researchers can identify autoantibodies with statistical significance and association with an infection or disease of interest. Generally, the patient sera are first diluted (e.g. 1,000-fold) before incubation on a pre-blocked antigen microarray (e.g. protein microarray) and ending with a stringent washing step. Then, positive signals are detected using anti-human IgG, IgM, or IgA antibodies conjugated to various fluorophores. Compared to traditional serology techniques, such as ELISA, agglutination, precipitation, complement-fixation, and fluorescent antibodies, protein microarray-based profiling is much more sensitive, unbiased, and can be performed at a much higher throughput. This next section will review four studies illustrating the history and development of protein microarrays in biomarker identification.

SARS-CoV diagnosis

The first viral proteome microarray was fabricated by Zhu et al. and consisted of every full-length protein and protein fragment encoded by the SARS coronavirus (SARS-CoV) as well as proteins from five additional mammalian coronaviruses (Zhu et al., 2006). These microarrays were used to screen 400 Canadian serum samples that were collected during the 2002 SARS outbreak. Included samples were those confirmed as SARS-CoV positive, other respiratory illness patients, and healthcare professionals. Antibody response was quantified by using both human IgM and IgG antibodies coupled

to different fluorophores. In order to identify biomarkers, the serum samples were first clustered according to the relative signal intensities of all coronavirus proteins in an unsupervised fashion. Two major groups were identified, which, when compared to clinical data, were largely correlated with either SARS-positive or SARS-negative sera. Five fragments of the SARS nucleocapsid protein (N protein) associated tightly with SARS infection, as well as one spike protein (S protein) fragment. However, a few proteins encoded by other coronaviruses also displayed significant correlation. In order to determine the best classifiers and classification model, two different supervised analysis approaches were applied; k nearest neighbor (k-NN), which measures the similarity between a new case and all known cases, and logistics regression (LR), a generalized linear regression for binary response. The N protein of SARS-CoV and the S protein from both SARS-CoV and HCoV-229E were identified as the best classifiers. One useful feature of a serum test over a nucleic acid diagnostic test is that anti-pathogen antibodies can potentially be detected long after infection. To that end, serum samples collected from SARS patients who recovered from respiratory disease (≈320 days after diagnosis) were used to probe the microarray, where positive signals were detected with both anti-human IgM and IgG antibodies. The results showed that SARS N proteins could be readily recognized by human IgG and importantly, not by IgM antibodies. However, serum samples collected from Chinese patients immediately after fever was detected showed much stronger signal both in IgG and IgM profiling. These results show that protein microarrays can be used to detect both early response anti-pathogen antibodies as well as late response antibodies long after infection. This approach is

potentially applicable to all viruses and is expected to have a large impact on both epidemiological studies and clinical diagnoses.

Humoral immune responses to herpesviruses

A similar approach has been used to profile humoral immune responses to two human herpesviruses, the Epstein-Barr virus (EBV) and Kaposi's sarcoma-associated herpesvirus (KSHV). While EBV is a ubiquitous human herpesvirus, KSHV has a much more restricted seroprevalence. Both viruses have been shown to be associated with malignancies and also display an increased frequency in individuals who are also infected with human immunodeficiency virus type 1 (HIV-1). To investigate humoral immune responses, a protein microarray consisting of 174 EBV and KSHV full-length proteins was generated by the Zhu and Hayward groups (Zhu et al., 2009; Zheng et al., 2011). Plasma antibody responses to EBV and KSHV were examined from healthy volunteers and patients with B-cell lymphoma, or with AIDS-related Kaposi's sarcoma or lymphoma. Apart from detecting IgG responses from known antigens, the tegument proteins ORF38 (KSHV), BBRF (EBV), BGLF2 (EBV), and BNRF1 (EBV), and the EBV early lytic proteins BRRF1 and BORF2, were also detected. IgA responses to EBV EBNA1 and viral capsid antigens have been used as a diagnostic tool for nasopharyngeal carcinoma for years, but the same IgA response was also found in healthy and HIV-infected patients. Comparing the IgG and IgA responses showed that IgA responses were much higher against BCRF1, BRRF2, and LMP2A. This study demonstrates that even

plasma can be used for biomarker identification and that other immunoglobulin isotypes such as IgA are worth considering when studying immune response.

E. coli proteome microarrays for IBD diagnosis

Crohn's disease (CD) and ulcerative colitis (UC) are chronic, idiopathic, and clinically heterogeneous intestinal disorders that are collectively known as inflammatory bowel disease (IBD). Although IBDs have been implicated in autoimmune disease, antibodies against microbes have been seen in the sera of IBD patients, some of which have been used as biomarkers for diagnosis and prognosis of the disease. Using the same *E. coli* K12 protein microarray that was used to study protein-DNA interactions mentioned earlier, Chen et al. decided to profile serum samples collected from CD and UC patients (Chen et al., 2009). The proteome array was screened using individual serum from healthy controls (n=39) and clinically well-characterized patients with CD (n=66) and UC (n=29). To their surprise, among the 417 *E. coli* proteins that were differentially recognized by serum antibodies from healthy controls and either IBD patient, 169 proteins were identified as highly immunogenic in healthy controls, 186 were identified as highly immunogenic in CD patients, but only 19 proteins were identified as highly immunogenic in UC patients. Through statistical analysis, they were able to identify two sets of serum antibodies as novel biomarkers for distinguishing CD from healthy controls (accuracy, 86±4%; $p < 0.01$) and CD from UC (accuracy, 80±2%; $p < 0.01$). This was the first demonstration of using proteome microarrays for the discovery

of novel serological markers, as well as the first to examine human immune responses to the entire proteome of a microbial species in a disease context.

Autoantigen discovery for Autoimmune Hepatitis (AIH)

A microarray consisting of individually purified human proteins would seem an ideal tool for the discovery of new autoantigens associated with autoimmune disease. Autoimmune hepatitis (AIH) is a chronic necroinflammatory disease of the human liver with little known etiology. Although the detection of non-organ-specific and liver-related autoantibodies using immunoserological approaches have been widely used for diagnosis and prognosis, these traditional autoantigens, such as anti-SMA (smooth muscle autoantibodies) and anti-ANA (antinuclear autoantibodies) are often mixtures of very complex biological materials. Identification and characterization of these autoantigens is dependent on their unambiguous and accurate detection, which is not possible using the traditional means. To address this, Song and colleagues created a human protein microarray of 5,011 non-redundant proteins that were expressed and purified from yeast (Song et al., 2010). In order to make this technique feasible in a clinical application, the cost of performing such an assay has to be considered, as a single human protein array with 9,000 proteins can cost upwards of $1,000. Therefore, a two-stage strategy was employed to limit costs in identifying new biomarkers in AIH. Phase I consists of rapid selection of candidate biomarkers, which are then validated in a smaller array in Phase II. In Phase I, 30 AIH and 30 control serum samples were obtained and individually used to probe the human protein microarrays at a 1,000-fold dilution. This was followed by

detection using a Cy5-conjugated anti-human IgG antibody. After statistical analysis, 11 candidate autoantigens were found. In order to validate the candidates, the 11 proteins and 3 positive controls were purified again to use in the production of a large number of low-cost small arrays for Phase II validation. These new, smaller arrays were sequentially probed with the serum samples used in Phase I as well as from serum samples obtained from an additional group of patients consisting of 22 AIH, 50 primary bilary cirrhosis (PBC), 43 hepatitis B (HB), 41 hepatitis C (HC), 11 system lupus erythematosus (SLE), 11 primary Sjögren's syndrome (pSS), and 2 rheumatoid arthritis (RA) samples. Twenty-six serum samples from patients with other types of severe disease and 50 samples from healthy subjects were used as negative controls. From these Phase II studies, three new antigens, RPS20, Alb2-like, and dUTPase, were identified as highly AIH-specific biomarkers with sensitivities of 47.5%, 45.5%, and 22.7%, respectively. These were further validated with additional AIH samples in a double-blind design, and were also able to demonstrate that these new biomarkers could be easily applied to canonical ELISA-based assays for clinical diagnosis and prognosis.

This study represents a novel paradigm in biomarker identification using protein microarrays for three reasons. First, a manageable number of candidate biomarkers can be quickly identified at a low cost because fewer expensive microarrays are required for Phase I. Second, by using a smaller microarray consisting only of selected candidate proteins, the validation step can be quickly carried out in higher throughput with a much lower cost. This is the crucial step in the validation of these markers if one is to avoid an "overfitting" problem that can arise when a candidate list is smaller than 40, which is likely the case for many of these applications. Overfitting becomes an issue when a

statistical model describes random error instead of identifying a relationship. If a system is excessively complex, as is the case with most biomarker identification screens, then there are too many individual-to-individual variations relative to the number of samples used, leading to poor predictive performance. Therefore, testing in an additional, larger cohort using a double-blind protocol is an effective way to overcome this issue. Third, the author developed an ELISA-based assay to examine the performance of the newly identified biomarkers, serving as a translational step towards clinical practice.

Several other studies have employed pathogen protein microarrays to profile serological responses following infection as well, including bacterial and viral arrays used for biomarker identification in various infectious diseases (Liang et al., 2011; Vigil et al., 2011; Luevano et al., 2010; Doolan et al., 2008). While these studies have clearly demonstrated the power of protein microarrays in the identification of potential biomarkers, several shortcomings are repeatedly seen in these studies. For example, many of these arrays were fabricated using proteins translated in *E. coli* lysates without purification. Because these proteins are contaminated with unwanted *E. coli* proteins, the sensitivity of the assay is likely reduced due to their high immunogenicity, although using *E. coli* lysates to block the array alleviates this problem to a degree. More importantly, many of the biomarkers identified in these studies were not validated with additional cohorts, therefore, the possibility of overfitting is not completely ruled out.

Pathogen-host Interactions

Another emerging application for protein microarrays in clinical proteomics is the unbiased, proteome-wide survey of important players involved in pathogen-host interactions. The identified factors, whether encoded by either a pathogen or a host, have the potential to be developed into potent therapeutic targets. This strategy is particularly useful for investigating virus-host interactions because after entering the host cells, the viral genome and its encoded proteins are in direct physical contact with the host's biological materials. This section will cover three ways in which these interactions can be investigated: RNA-protein interactions, enzyme-substrate relationships, and protein-protein interactions.

BMV RNA and Host Proteome Interactions

The yeast proteome array was used to identify host factors that can affect the replication of Brome Mosaic Virus (BMV), an RNA virus that infects plants, but can also replicate inside of *S. cerevisiae* (Zhu et al, 2007). Previous studies had shown that this positive-stranded RNA virus encodes a tRNA-like structure at the 3'-end of its RNA genome, where a clamped adenine motif (CAM) is required for the packaging of its genome into the capsid. To identify the necessary host proteins that could interfere with the viral packaging process, the yeast proteome arrays were incubated with a Cy3-labeled CAM-containing RNA stem-loop structure in the presence of a Cy5-labeled mutated CAM hairpin. By taking the Cy3:Cy5 ratio, the top hits were identified and validated

using an *in vitro* gel-shift assay. Two candidate proteins, Pseudouridine Synthase 4 (Pus4) and the Actin Patch Protein 1 (App1), were further characterized in *Nicotiana benthamiana*. Both were shown to reduce BMV plus-strand RNA accumulation modestly, but had a much greater effect on the ability of BMV to spread in plants. Pus4 was also shown to prevent the encapsidation of the BMV RNA and reassembly of virions, providing a regulatory link between CAM-containing RNA motifs and encapsidation. This is yet another example of the considerable versatility of protein microarrays. Not only can they be used to study prokaryotic and eukaryotic interactions, but viral interactions as well, in an unbiased and systematic manner.

Host Phosphorylome of Virus-Encoded Kinases

Viruses have been very successful in exploiting their hosts in developing their own "arsenal," sometime using both host DNA and proteins in the process. Understanding the method by which the virus interacts with the host machinery is critical in developing effective antivirals. The human α, β, and γ herpesviruses differently infect tissues and cause distinct diseases, ranging from mild cold sores to pneumonitis, birth defects, and even cancers. However, they each confront similar challenges in infecting their hosts, namely reprogramming cellular gene expression, sensing and modifying cell-cycle progression, and reactivating the lytic life cycle to produce new virions and spread infection. Across the herpesvirus family, many orthologous serine/threonine kinases are conserved that are involved in replication of the viral genome during the lytic life cycle. Therefore, many have postulated that if there are any shared substrates between these

orthologous viral kinases, the targeted host proteins critical for viral replication would be revealed and aid in distinct and effective antiviral targets.

In testing this hypothesis, Li and colleagues purified four orthologous kinases encoded by EBV, KSHV, HCMV (human cytomegalovirus), and HSV-1 (herpes simplex virus 1) and incubated them on a human protein microarray (Li and Zhu et al., 2011). They identified 110 shared substrates and applied Gene Ontology (GO) and STRING analyses (a database of known and predicted protein-protein interactions – http://string-db.org/) to these candidates, finding a highly connected cluster of 15 proteins. Surprisingly, these proteins were all shown to be associated with the DNA damage response (DDR) pathway. The host DDR has been known to be an important player for many viruses, including human herpesviruses, and is also relevant to virus-induced tumorigenesis (Nikitin et al., 2012). To further narrow down the potential players in this pathway, it was reasoned that an upstream master regulator of the pathway would be a likely target. After literature searches, Tat-interactive protein 60 (TIP60) emerged as an excellent candidate for four reasons: (1) TIP60 is further upstream in the DDR pathway than the other candidates; (2) it serves as a master regulator in DDR by acetylating ATM and initiating its autophosphorylation activity; (3) it is a histone acetyltransferase, and therefore can regulate chromatin dynamics through histone acetylation; and (4) it has been implicated in other viruses. When TIP60 was knocked down in EBV-infected B-cells, EBV's lytic replication was greatly reduced. Next, the authors performed a series of cell-based assays showing that during EBV replication, TIP60 activation by the BGLF4 kinase triggers EBV-induced DDR and mediates induction of viral lytic gene

expression. Efficient lytic replication was also shown to be dependent on TIP60 in KSHV, HCMV, and HSV-1.

This work illustrates that high-throughput, unbiased approaches for the identification of conserved viral host targets show great potential for novel therapeutic antiviral targets. Treatment of herpesvirus is limited to very few drugs due to viral escape mutants that arise from extensive use, rendering these kinases attractive targets for future studies. Like all drug studies, development of effective drugs requires extensive knowledge of cellular targets. This novel "common substrate identification" approach using microarrays can aid in the design of new assays for new and broadly effective anti-herpesvirus therapeutics.

LANA Interactome Analysis Reveals a Role in Telomere Shortening

While we have already discussed using protein microarrays to measure protein-protein interactions, it can also be specifically used to profile interactomes between a pathogen and a host. Recently, Shamay and colleagues examined interactions between the KSHV-encoded virulent factor, LANA, and a human host using the transcription factor microarray described before (Shamay et al., 2012). LANA is an essential participant in KSHV genome replication and dysregulated cell growth in latently infected cells. Its interactions with host proteins have been extensively studied using a variety of techniques (including Y2H, GST affinity, IP assays, and HPLC-MS/MS), but each approach identified a different group of proteins. This study used a FLAG-tagged LANA to identify 61 potential binding partners, many of which were novel. Further validations

confirmed 8 of 9 by co-immunoprecipitation and included TIP60, protein phosphatase 2A (PP2A), replication protein A (RPA) and the DNA-repair protein XPA. Although human papillomavirus (HPV) E6, HIV-1 TAT, and HCMV pUL27 interact with TIP60 to induce its degradation, LANA-associated TIP60 retained its HAT activity and showed increased stability. This finding mirrors what was described in the previous section that showed TIP60 plays a positive role in KSHV lytic replication. The identification of LANA as an RPA binding partner suggests that it may also have a role in the replication of cellular telomeric DNA. To test this, the authors performed ChIP assays using anti-RPA1 and –RPA2 antibodies and primers specific for the telomeric regions of DNA. They found that when LANA was present, RPA1 and RPA2 recruitment to telomeres was inhibited, but had no impact on the protein levels of the RPA complex, which implicated LANA as a factor in telomere length. Using Southern blots to analyze terminal restriction fragments showed that the length of telomeres was shortened by at least 50% in both LANA-expressing endothelial cells and KSHV-infected primary effusion lymphoma cells. Although it was shown that LANA effects telomere length, many more questions regarding its role need to be answered to fully understand how it is used by KSHV.

SUMO-EBV Interactome Reveals a New Mechanism of EBV Lytic Replication

Inversely, a human factor of interest can be used to probe a virus protein microarray to identify crucial viral factors. The small ubiquitin-related modifier (SUMO) pathway is similar to the ubiquitin pathway because it involves a series of sequential enzymatic reactions to conjugate SUMO to lysine residues on substrate proteins. It has

been shown that both latent and lytic EBV proteins can interact with the SUMOylation pathway, with non-covalent interactions occurring between SUMO and EBV proteins via a SUMO interaction motif (SIM). In order to systematically identify EBV proteins that bind to SUMO, Li et al performed binding assays with human SUMO2 using the previously described EBV proteome microarray (Li et al., 2012). Eleven proteins were identified, including the conserved viral kinase BGLF4. Mutations at putative SIMs within BGLF4 at both the N- and C-termini changed the intracellular localization from the nucleus to the cytoplasm. If mutations were made at only the N-termini, localization remained mostly in the nucleus, and localization was in both the cytoplasm and the nucleus when mutations were made at only the C-termini. BGLF4 was shown to inhibit the EBV lytic cycle transactivator ZTA by abolishing its SUMOylation. The inhibition was shown to be dependent on both BGLF4 SUMO binding and BGLF4 kinase activity. Global SUMOylation was suppressed by active BGLF4, but not by SIM or kinase-dead BGLF4 mutants. Furthermore, interaction of BGLF4 with SUMO was required to induce DDR and enhance extracellular viral production during EBV lytic replication.

Identification of Novel Streptococcal Proteins that Bind Human Ligands

Pathogen-host interactions are not just limited to viruses, and the need to find novel antibiotics is becoming increasingly pertinent. Therefore, it is not surprising that protein microarrays have been utilized to identify bacterial pathogen-host interactions. Margarit and colleagues used microarrays to identify proteins expressed by two species of the streptococcus gram-positive bacteria, *Streptococcus pyogenes* and *S. agalactiae,*

that interact with human factors known to mediate pathogenesis (Margarit et al., 2009). A bioinformatics approach was first used to predict 200 proteins that are present on the cellular surface, and therefore are more likely to paly a role in infection. The human probes were also carefully selected because of their known roles in colonization and infection: fibronectin, fibrinogen, and C4 binding protein. Treating the array consisting of the 200 predicted proteins with these probes identified 17 of the 20 known interactions, as well as 8 novel contacts, which were further validated using far-western blot analysis. Three of the novel proteins identified were related, termed fib proteins, and using domain mapping, the authors were able to identify regions of these proteins that were required for their interaction with the human ligands. Subsequent assays of sera from patients with *S. agalactiae* showed high titers of Fib-specific antibodies, indicating their relative abundance during infection. There is hope that future studies will be able to determine the exact role of these proteins in infection and if they are suitable drug targets for Streptococcus infections.

Taken together, these studies show the tremendous power of protein microarrays in the discovery of novel molecular mechanisms underlying host-pathogen interactions.

Disease Type	Disease	Substrate	Reference
Infectious	SARS infection	82 viral proteins	Zhu et al., 2006
	B cell lymphoma/ AIDs-related Kaposi's lymphoma	174 EBV and KSHV viral proteins	Zheng et al., 2011
	Rabbit model of the plague	149 proteins from *Yersinia pestis*	Li et al., 2005
	Brucellosis	3,046 proteins expressed in lysates	Liang et al., 2011
	Cervical carcinomas	154 proteins from 13 papillomavirus	Luevano et al., 2010
	Streptococcus infection	201 viral proteins from 2 strains	Margarit et al., 2009
Autoimmune	Inflammatory bowel disease	*E. coli* arrays with 4,179 proteins	Chen et al., 2009
	Autoimmune hepatitis	5,011 human proteins	Song et al., 2010
	Primary biliary cirrhosis	17,000 human proteins	Hu et al., 2012
	Sjögren's syndrome	8,000 human proteins	Hu et al., 2011

Table 5. Summary of Clinical Disease Studies Using Protein Microarrays

Cancer Research

Over the last five years, rapid development of genome-wide sequencing (e.g. next-gen sequencing) has revealed the heterogeneous nature of tumors, but clinical diagnosis is largely still dependent on morphological patterns (Parsons et al., 2011; Gerlinger et al., 2012). Although tumors have indistinguishable morphology, they often have vastly different clinical outcomes, so the composition of this heterogeneity needs to be better understood before more effective therapies can be developed. An individualistic therapy approach is thought to be the future of cancer research, so a new class of proteomic profiling technologies will have to be developed in order to meet this need. Protein microarrays have shown the potential to meet this need and have proven useful in profiling the functional state of tumors and for cancer biomarker identification.

Reverse Phase Protein Microarrays in Cancer Biology

One common approach that has been used to determine the status of signaling pathways in tumor cells is based on an immunoblot assay that uses antibodies to recognize phosphorylated proteins. Haab et al was the first to scale this up to high throughput when they developed an antibody array, spotting individual commercial antibodies onto glass slides in high density (Habb et al., 2001). This technology allows for the simultaneous detection of multiple antigens presented in a complex biological sample, such as cells, tissues, and body fluids (Borrebaeck et al., 2009; Haab, 2003).

In 2001, Paweletz and colleagues coined the term "reverse phase protein microarray" to describe an array in which lysates of cells or tissues are immobilized on the array surface rather than antibodies (Paweletz et al., 2001). By using phosphoprotein-specific antibodies, these arrays can be used to interrogate the phosphorylation state present in cell lysates from tumor samples. Many clinical trials are currently ongoing that are utilizing these arrays (Muellet et al., 2010). A large hurdle for most clinical trials and diagnostic tests is the minimal amount of sample that is available from the patients. Unobtrusive methods are few and far between, so these arrays are a welcome approach because not only can it allow for the evaluation of multiple components of a signaling pathway, very little sample is required for microarray fabrication. These samples can be printed on a series of identical arrays and analyzed in parallel using commercially available phoshoprotein-specific antibodies.

Petricoin and coworkers obtained 59 cell lysate samples from the Children's Oncology Group Intergroup Rhabdomyosarcoma Study (IRS), and fabricated them into a reverse phase microarray (Petricoin et al., 2007). Rhabdomyosarcoma is a rare childhood cancer that arises from undifferentiated muscle progenitor cells. Current treatment can yield a disease-free survival rate of close to 67%, but the reason for the failure of the treatment for the remaining one-third of patients is largely misunderstood. Identification of biomarkers could help distinguish these patients from the responders to the traditional therapy, which would aid in pursuing better therapies, and possibly identify other novel potential drug targets. The reverse phase microarrays were used as a platform to detect the phosphorylation status of proteins (using phospho-specific antibodies) thought to be distinguishing factors between the rhabdomyosarcoma

subtypes. The group was able to identify higher phosphorylation levels in 4 Akt/mTOR pathway components that were only in patients with poor survival outcomes. Conversely, through bioinformatics analysis, they found that patients with good treatment outcomes exhibited mTOR pathway suppression. Together, these findings suggest that pharmacologically suppressing the mTOR pathway could result in improved outcomes for patients who did not respond to the standard chemotherapy treatment. Using a mouse xenograft model of rhabdomyosarcoma cell lines, there were able to test this hypothesis with known mTOR pathway inhibitors. As expected, treatment with the inhibitor resulted in reduced phosphorylation of the protein 4E-BP1, a protein that had been identified in the array screen, as well as reduced tumor growth. As a further proof of principle, they were able to demonstrate with the cell lysate arrays that specific patient sub-populations could be identified that benefited from personalized therapies and were able to identify the specific molecules that should be targeted in order to test these therapies.

Although this technique holds promise, there are several major problems with this approach. First, well-characterized antibodies are not available for a great majority of human proteins, especially phosphoprotein-specific ones (Kalyuzhny, 2009; Couchman, 2009). Second, recent studies have advocated that many of the commercially available monoclonal antibodies may not even recognize their claimed targets, and can cross-react extensively with other cellular components (Jensen et al., 2009). Third, when using antibodies for diagnostic and therapeutic applications, cross-reactivity is a large concern, as indicated by the recent withdrawal of several mAb-based pharmaceuticals from the market (Hughes, 2009; Berger et al., 2009). Finally, this technique is not ideal for the

discovery of novel biomarkers, as prior knowledge and phospho-specific antibodies are required for any evaluation.

Identification of Autoantibody Biomarkers for the Early Detection of Breast Cancer

While breast cancer awareness has resulted in many saved lives, current screening using mammograms detects only 70% of breast cancers, and false-positive mammograms leads to painful and unnecessary biopsies. If one were to identify biomarkers allowing for early detection of breast cancer, it could provide a non-invasive and low cost method that also improves patient outcomes. One arising category of cancer biomarkers is autoantibodies to tumor antigens. They offer better stability, specificity, ease of purification and detection compared to other serum proteins. Anderson et al treated NAPPA protein arrays (described previously) containing tumor antigens with either breast cancer patient or control serum samples, in order to identify any differences in the human antibody repertoire that could be used as a biomarker (Anderson et al., 2011). The first step in identifying new biomarkers used arrays containing 4,988 tumor antigens in order to eliminate uninformative autoantibodies that were present in both patient and control samples. After identifying these background antigens, the number was reduced to 761, which allowed them to fabricate much smaller arrays for the next phase of study that were cheaper and would contain less false positives. In this next phase, sera from patients with invasive early breast cancer and benign breast disease were compared, resulting in 119 antigens that were present only in invasive cancer. The third phase validated 28 antigens that maintained high levels of specificity in a blinded assay, one of

which was a known autoantigen ATP6AP1. They were able to show high expression of ATP6AP1 in four breast cancer cell lines by western blot as well as higher ATP6AP1 autoantibody levels in approximately 13% of early breast cancer serum samples compared to control. While only a first step, this work demonstrates the usefulness of protein microarrays, particularly NAPPA arrays, in the identification of early biomarkers in breast cancer.

Finding Autoantibody Biomarkers in Bladder Cancer

The utility of cancer biomarkers is in the development of new strategies for early diagnosis of disease, which allows for early intervention with current therapies and the improvement of patient survival rates. Cancer-associated autoantibodies often target proteins that are mutated, modified, or aberrantly expressed in tumor cells, so they can be useful immunologic reporters that identify the molecular mechanisms underlying tumorigenesis. This, in turn, may represent a good starting point for the development of new treatments. With regards to bladder cancer, Orenes-Piñero and colleagues sought to identify not only autoantibody biomarkers of early bladder cancer, but also understand the underlying pathology of the disease (Orenes-Piñero et al., 2010). Comparing serum samples collected from 12 patients with bladder cancer with 10 control patients using the Invitrogen Protoarray (an array consisting of ≈8,000 purified human proteins), they identified 171 differentially expressed proteins.

Two proteins, clusterin and dynamin, were among the identified and were further validated using a custom tissue microarray comprised of bladder cancer tumor samples.

They applied immunohistochemistry to these arrays and found reduced expression levels of clusterin in muscle invasive bladder cancer, compared to non-muscle invasive tumors. It was also shown that low expression of dynamin was associated with increased tumor stage and grade, higher recurrence rate after surgery, and shorter survival. Subsequent findings associating low levels of both clusterin and dynamin with disease appeared contradictory to the initial observation, which showed an increase in autoantibody levels to these proteins in patients with bladder cancer. However, despite these contrary findings, they were able to demonstrate significant associations between clusterin and dynamin protein levels and disease progression, and still potentially make use of them as biomarkers in the clinic and for drug screens.

Cancer Type	Substrate	Probe Type	Findings	Reference
Bladder	8,000 human proteins	Patient sera	171 serum autoantibodies	Orenes-Pinero et al., 2010
	254 serum protein antibodies	Patient sera	Protein profiles (including C-Met)	Sanchez-Carbayo et al., 2006
Breast	4,988 candidate tumor antigens	Patient sera	119 serum autoantibodies	Anderson et al., 2011
	94 lectins	Mammalian cell lines	Cancer stem-cell like glycan signatures	Tao et al., 2008
	Cytokine antibody array	MGF7/Her2-18 cell line	Cytokine signatures from breast cancer cell line	Vazquez-Martin et al., 2007
	378 antibody array	Malignant breast tissue	Cancer expression profiles	Hudelist et al., 2004
Lung	378 antibody array	Patient tissue samples	29 different-ially regulated targets	Bartling et al., 2005
Pancreatic	48 antibody array	Patient sera	Glycan-alterations of MUC1 and CEA	Chen et al., 2007
Prostate	184 antibody array	Patient sera	5 differentially regulated targets	Miller et al., 2003
	107 antibody array	Patient sera	Several targets, including Trombospondin-1	Shafer et al., 2007
Rhabdomyo-sarcoma	59 cancer cell lysates	Phospho-specific antibodies	Phosphorylation status of 27 proteins	Petricoin et al., 2007

Table 6. Summary of Cancer Studies Using Protein Microarrays

1.7 Data Analysis and Bioinformatics

One common characteristic of all microarray experiments is the generation of large quantities of data. In order to elucidate interactions, careful statistical analysis is often required. This is most commonly performed by bioinformaticians, who work in close collaboration with laboratory scientists to make sense of the vast pools of data generated from a typical experiment. The analysis of data generated from protein microarray studies can be divided into two stages, or levels. First level analysis involves obtaining a reliable signal intensity value for each spot on a protein microarray for a particular assay. Second level analysis is more advanced analysis built upon the profile of signal intensity values obtained from the first level analysis.

<u>First level analysis</u>

The analysis at this level includes data acquisition, data normalization, and identification of hits. The raw data directly acquired from protein microarrays contain noise, which arise from various sources, such as inconsistent printing, uneven distribution of the reaction mixtures, inconsistent washing conditions, and batch effects. In most cases, algorithms developed for DNA/oligo microarray analysis can be applied to protein microarray analysis (Smyth et al., 2003). It is often the case that proteins in a local region of the microarray show higher signal intensity than the overall intensity of the entire array. Since proteins are usually printed in a random order on a protein microarray, without regard for biological structure or function, it is not expected that such

an effect is biological. To correct for this noise, the most useful normalization is the local normalization algorithm (e.g., ProCAT; Zhu et al., 2006). For each spot, the average intensity of a small set of proteins (e.g., a 3-by-3 grid with the spot of interest in the middle) surrounding the spot as the local background intensity is used. Next, the local background is subtracted from the intensity of the spot of interest to achieve local noise correction. This process is then repeated for all spots on a protein microarray.

The next step is to identify positive hits for a particular assay. Because of the high cost associated with protein microarrays, replicates are often not employed, so a vigorous significance (e.g., *P* value) for each protein can generally not be calculated. Rather, the overall intensity distribution of proteins on a protein microarray is calculated. In many cases, the distribution can be fit by a normal distribution curve. Proteins with intensity above a certain number of standard deviations (e.g. 3 SDs) are considered positive hits. In some situations, a set of negative control experiments can be performed, generating intensity profiles that can be used to estimate the false discovery rate.

Second level analysis

Various bioinformatics tools are utilized or developed for the second level analyses. Using biomarker identification as an example, the main question asked involves how to classify the groups of clustered proteins in order to draw conclusions. In other words, the computational problem is identifying an optimal set of proteins that can best separate the disease from the control samples. Normally, if one can identify elevated

autoantibodies in disease samples relative to the controls, the corresponding autoantigens (i.e., human proteins) can serve as biomarkers for the disease.

There are two major steps of the classification: 1) selection of a set of proteins, and 2) evaluation of the performance of this combination of these proteins. Simply using the most antigenic protein may not be meaningful because its specificity and sensitivity is often not sufficient for diagnosis. In addition, a brutal force search for protein combinations is also not feasible. Different machine learning algorithms can be used to select the best combinations. Examples of such models include those based on multivariate logistic regression, mixtures of Gaussian distributions (Skates et al., 2004), decision trees, artificial neural networks (ANNs), support vector machines (SVMs), and various approaches to incorporate these models in some kind of ensembles to improve overall performance (Zhang et al., 2006; Zhang et al, 2010). In addition, heuristic searching algorithms, such as sequential forward search and sequential forward floating search, can also be used to search for the best combination of features (i.e., proteins) (Sahiner et al, 2000).

The performance will be assessed by sensitivity, specificity, and receiver-operating characteristic (ROC). To derive models that are robust and hence, more likely biologically based, and to alleviate the constraint of typical suboptimal sample sizes, statistical re-sampling methods, such as bootstrap, k-fold or leave-one-out cross-validation and introduction of artificial perturbations, have been widely utilized. These combined with known databases and gene ontology (GO) analysis, narrow down hits to more biologically relevant targets that can be further characterized in the laboratory.

1.8 References

Anderson, K.S., Sibani, S., Wallstrom, G., Qiu, J., Mendoza, E.A., Raphael, J., Hainsworth, E., Montor, W.R., Wong, J., Park, J.G.*, et al.* (2011). Protein microarray signature of autoantibody biomarkers for the early detection of breast cancer. J. Proteome Res. *10,* 85-96.

Angenendt, P., Kreutzberger, J., Glokler, J., and Hoheisel, J.D. (2006). Generation of high density protein microarrays by cell-free in situ expression of unpurified PCR products. Mol. Cell. Proteomics *5,* 1658-1666.

Barbulovic-Nad, I., Lucente, M., Sun, Y., Zhang, M., Wheeler, A.R., and Bussmann, M. (2006). Bio-microarray fabrication techniques--a review. Crit. Rev. Biotechnol. *26,* 237-259.

Bartling, B., Hofmann, H.S., Boettger, T., Hansen, G., Burdach, S., Silber, R.E., and Simm, A. (2005). Comparative application of antibody and gene array for expression profiling in human squamous cell lung carcinoma. Lung Cancer *49,* 145-154.

Berger, J.R., Houff, S.A., and Major, E.O. (2009). Monoclonal antibodies and progressive multifocal leukoencephalopathy. Mabs *1,* 583-589.

Borrebaeck, C.A., and Wingren, C. (2009). Design of high-density antibody microarrays for disease proteomics: key technological issues. J. Proteomics *72,* 928-935.

Brichta, J., Hnilova, M., and and Viskovic, T. (2005). Generation of hapten-specific recombinant antibodies: Antibody phage display technology: A review. Vet. Med. *50,* 231-252.

Bussow, K., Cahill, D., Nietfeld, W., Bancroft, D., Scherzinger, E., Lehrach, H., and Walter, G. (1998). A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. Nucleic Acids Res. *26,* 5007-5008.

Carmen, S., and Jermutus, L. (2002). Concepts in antibody phage display. Brief Funct. Genomic Proteomic *1,* 189-203.

Chang, T.W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. J. Immunol. Methods *65,* 217-223.

Chao, G., Lau, W.L., Hackel, B.J., Sazinsky, S.L., Lippow, S.M., and Wittrup, K.D. (2006). Isolating and engineering human antibodies using yeast surface display. Nat. Protoc. *1,* 755-768.

Chen, C.S., Korobkova, E., Chen, H., Zhu, J., Jian, X., Tao, S.C., He, C., and Zhu, H. (2008). A proteome chip approach reveals new DNA damage recognition activities in Escherichia coli. Nat. Methods *5,* 69-74.

Chen, C.S., Sullivan, S., Anderson, T., Tan, A.C., Alex, P.J., Brant, S.R., Cuffari, C., Bayless, T.M., Talor, M.V., Burek, C.L.*, et al.* (2009). Identification of novel serological biomarkers for inflammatory bowel disease using Escherichia coli proteome chip. Mol. Cell. Proteomics *8,* 1765-1776.

Chen, C.S., and Zhu, H. (2006). Protein microarrays. Biotechniques *40,* 423, 425, 427 passim.

Chen, S., LaRoche, T., Hamelinck, D., Bergsma, D., Brenner, D., Simeone, D., Brand, R.E., and Haab, B.B. (2007). Multiplexed analysis of glycan variation on native proteins captured by antibody microarrays. Nat. Methods *4,* 437-444.

Couchman, J.R. (2009). Commercial antibodies: the good, bad, and really ugly. J. Histochem. Cytochem. *57,* 7-8.

Diamond, D.L., Zhang, Y., Gaiger, A., Smithgall, M., Vedvick, T.S., and Carter, D. (2003). Use of ProteinChip array surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) to identify thymosin beta-4, a differentially secreted protein from lymphoblastoid cell lines. J. Am. Soc. Mass Spectrom. *14,* 760-765.

Dias, W.B., Cheung, W.D., and Hart, G.W. (2012). O-GlcNAcylation of kinases. Biochem. Biophys. Res. Commun. *422,* 224-228.

Doolan, D.L., Mu, Y., Unal, B., Sundaresh, S., Hirst, S., Valdez, C., Randall, A., Molina, D., Liang, X., Freilich, D.A.*, et al.* (2008). Profiling humoral immune responses to P. falciparum infection with protein microarrays. Proteomics *8,* 4680-4694.

Ekins, R.P. (1989). Multi-analyte immunoassay. J. Pharm. Biomed. Anal. *7,* 155-168.

Evans-Nguyen, K.M., Tao, S.C., Zhu, H., and Cotter, R.J. (2008). Protein arrays on patterned porous gold substrates interrogated with mass spectrometry: detection of peptides in plasma. Anal. Chem. *80,* 1448-1458.

Festa, F., Steel, J., Bian, X., and Labaer, J. (2013). High-throughput cloning and expression library creation for functional proteomics. Proteomics

Foster, M.W., Forrester, M.T., and Stamler, J.S. (2009). A protein microarray-based analysis of S-nitrosylation. Proc. Natl. Acad. Sci. U. S. A. *106,* 18948-18953.

Fu, D., Ma, J., and Chen, J. (2007). Dual-wavelength microarray fluorescence detection system using volume holographic filter. J. Biomed. Opt. *12,* 014040.

Gavin, I.M., Kukhtin, A., Glesne, D., Schabacker, D., and Chandler, D.P. (2005). Analysis of protein interaction and function with a 3-dimensional MALDI-MS protein array. Biotechniques *39,* 99-107.

Ge, H. (2000). UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. Nucleic Acids Res. *28,* e3.

Gupta, R., Kus, B., Fladd, C., Wasmuth, J., Tonikian, R., Sidhu, S., Krogan, N.J., Parkinson, J., and Rotin, D. (2007). Ubiquitination screen using protein microarrays for comprehensive identification of Rsp5 substrates in yeast. Mol. Syst. Biol. *3,* 116.

Gutmann, O., Kuehlewein, R., Reinbold, S., Niekrawietz, R., Steinert, C.P., de Heij, B., Zengerle, R., and Daub, M. (2004). A highly parallel nanoliter dispenser for microarray fabrication. Biomed. Microdevices *6,* 131-137.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.*, et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458,* 223-227.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19,* 1720-1730.

Haab, B.B. (2003). Methods and applications of antibody microarrays in cancer research. Proteomics *3,* 2116-2122.

Haab, B.B., Dunham, M.J., and Brown, P.O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biol. *2,* RESEARCH0004.

Hall, D.A., Zhu, H., Zhu, X., Royce, T., Gerstein, M., and Snyder, M. (2004). Regulation of gene expression by a metabolic enzyme. Science *306,* 482-484.

He, M., and Taussig, M.J. (2001). Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). Nucleic Acids Res. *29,* E73-3.

Ho, S.W., Jona, G., Chen, C.T., Johnston, M., and Snyder, M. (2006). Linking DNA-binding proteins to their recognition sequences by using protein microarrays. Proc. Natl. Acad. Sci. U. S. A. *103,* 9940-9945.

Hu, C.J., Song, G., Huang, W., Liu, G.Z., Deng, C.W., Zeng, H.P., Wang, L., Zhang, F.C., Zhang, X., Jeong, J.S.*, et al.* (2012). Identification of new autoantigens for primary biliary cirrhosis using human proteome microarrays. Mol. Cell. Proteomics *11,* 669-680.

Hu, J., Rho, H.S., Newman, R.H., Hwang, W., Neiswinger, J., Zhu, H., Zhang, J., and Qian, J. (2013). Global analysis of phosphorylation networks in humans. Biochim. Biophys. Acta

Hu, S., Li, Y., Liu, G., Song, Q., Wang, L., Han, Y., Zhang, Y., Song, Y., Yao, X., Tao, Y.*, et al.* (2007). A protein chip approach for high-throughput antigen identification and characterization. Proteomics *7,* 2151-2161.

Hu, S., Vissink, A., Arellano, M., Roozendaal, C., Zhou, H., Kallenberg, C.G., and Wong, D.T. (2011). Identification of autoantibody biomarkers for primary Sjogren's syndrome using protein microarrays. Proteomics *11,* 1499-1507.

Hu, S., Xie, Z., Onishi, A., Yu, X., Jiang, L., Lin, J., Rho, H.S., Woodard, C., Wang, H., Jeong, J.S.*, et al.* (2009). Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. Cell *139,* 610-622.

Hu, S., Xie, Z., Qian, J., Blackshaw, S., and Zhu, H. (2011). Functional protein microarray technology. Wiley Interdiscip. Rev. Syst. Biol. Med. *3,* 255-268.

Huang, J., Zhu, H., Haggarty, S.J., Spring, D.R., Hwang, H., Jin, F., Snyder, M., and Schreiber, S.L. (2004). Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips. Proc. Natl. Acad. Sci. U. S. A. *101,* 16594-16599.

Hudelist, G., Pacher-Zavisin, M., Singer, C.F., Holper, T., Kubista, E., Schreiber, M., Manavi, M., Bilban, M., and Czerwenka, K. (2004). Use of high-throughput protein array for profiling of differentially expressed proteins in normal and malignant breast tissue. Breast Cancer Res. Treat. *86,* 281-291.

Hughes, B. (2010). Antibody-drug conjugates for cancer: poised to deliver? Nat. Rev. Drug Discov. *9,* 665-667.

Jensen, B.C., Swigart, P.M., and Simpson, P.C. (2009). Ten commercial antibodies for alpha-1-adrenergic receptor subtypes are nonspecific. Naunyn Schmiedebergs Arch. Pharmacol. *379,* 409-412.

Jeong, J.S., Jiang, L., Albino, E., Marrero, J., Rho, H.S., Hu, J., Hu, S., Vera, C., Bayron-Poueymiroy, D., Rivera-Pacheco, Z.A.*, et al.* (2012). Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. Mol. Cell. Proteomics *11,* O111.016253.

Jones, R.B., Gordus, A., Krall, J.A., and MacBeath, G. (2006). A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature *439,* 168-174.

Joung, J.K., Ramm, E.I., and Pabo, C.O. (2000). A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc. Natl. Acad. Sci. U. S. A. *97,* 7382-7387.

Kalyuzhny, A.E. (2009). The dark side of the immunohistochemical moon: industry. J. Histochem. Cytochem. *57,* 1099-1101.

Kodoyianni, V. (2011). Label-free analysis of biomolecular interactions using SPR imaging. Biotechniques *50,* 32-40.

Kopf, E., and Zharhary, D. (2007). Antibody arrays--an emerging tool in cancer proteomics. Int. J. Biochem. Cell Biol. *39,* 1305-1317.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P.*, et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440,* 637-643.

Kung, L.A., Tao, S.C., Qian, J., Smith, M.G., Snyder, M., and Zhu, H. (2009). Global analysis of the glycoproteome in Saccharomyces cerevisiae reveals new roles for protein glycosylation in eukaryotes. Mol. Syst. Biol. *5,* 308.

Kuno, A., Kato, Y., Matsuda, A., Kaneko, M.K., Ito, H., Amano, K., Chiba, Y., Narimatsu, H., and Hirabayashi, J. (2009). Focused differential glycan analysis with the platform antibody-assisted lectin profiling for glycan-related biomarker verification. Mol. Cell. Proteomics *8,* 99-108.

Landry, J.P., Sun, Y.S., Guo, X.W., and Zhu, X.D. (2008). Protein reactions with surface-bound molecular targets detected by oblique-incidence reflectivity difference microscopes. Appl. Opt. *47,* 3275-3288.

Lange, S.A., Benes, V., Kern, D.P., Horber, J.K., and Bernard, A. (2004). Microcontact printing of DNA molecules. Anal. Chem. *76,* 1641-1647.

Lee, J.R., Magee, D.M., Gaster, R.S., LaBaer, J., and Wang, S.X. (2013). Emerging protein array technologies for proteomics. Expert Rev. Proteomics *10,* 65-75.

Li, B., Jiang, L., Song, Q., Yang, J., Chen, Z., Guo, Z., Zhou, D., Du, Z., Song, Y., Wang, J.*, et al.* (2005). Protein microarray for profiling antibody responses to Yersinia pestis live vaccine. Infect. Immun. *73,* 3734-3739.

Li, R., Wang, L., Liao, G., Guzzo, C.M., Matunis, M.J., Zhu, H., and Hayward, S.D. (2012). SUMO binding by the Epstein-Barr virus protein kinase BGLF4 is crucial for BGLF4 function. J. Virol. *86,* 5412-5421.

Li, R., Zhu, J., Xie, Z., Liao, G., Liu, J., Chen, M.R., Hu, S., Woodard, C., Lin, J., Taverna, S.D.*, et al.* (2011). Conserved herpesvirus kinases target the DNA damage

response pathway and TIP60 histone acetyltransferase to promote virus replication. Cell. Host Microbe *10,* 390-400.

Li, Y., Tao, S.C., Bova, G.S., Liu, A.Y., Chan, D.W., Zhu, H., and Zhang, H. (2011). Detection and verification of glycosylation patterns of glycoproteins from clinical specimens using lectin microarrays and lectin-based immunosorbent assays. Anal. Chem. *83,* 8509-8516.

Liang, L., Tan, X., Juarez, S., Villaverde, H., Pablo, J., Nakajima-Sasaki, R., Gotuzzo, E., Saito, M., Hermanson, G., Molina, D.*, et al.* (2011). Systems biology approach predicts antibody signature associated with Brucella melitensis infection in humans. J. Proteome Res. *10,* 4813-4824.

Lin, Y.Y., Lu, J.Y., Zhang, J., Walter, W., Dang, W., Wan, J., Tao, S.C., Qian, J., Zhao, Y., Boeke, J.D., Berger, S.L., and Zhu, H. (2009). Protein acetylation microarray reveals that NuA4 controls key metabolic target regulating gluconeogenesis. Cell *136,* 1073-1084.

Lu, J.Y., Lin, Y.Y., Boeke, J.D., and Zhu, H. (2013). Using functional proteome microarrays to study protein lysine acetylation. Methods Mol. Biol. *981,* 151-165.

Lu, J.Y., Lin, Y.Y., Qian, J., Tao, S.C., Zhu, J., Pickart, C., and Zhu, H. (2008). Functional dissection of a HECT ubiquitin E3 ligase. Mol. Cell. Proteomics *7,* 35-45.

Lu, K.Y., Tao, S.C., Yang, T.C., Ho, Y.H., Lee, C.H., Lin, C.C., Juan, H.F., Huang, H.C., Yang, C.Y., Chen, M.S.*, et al.* (2012). Profiling lipid-protein interactions using nonquenched fluorescent liposomal nanovesicles and proteome microarrays. Mol. Cell. Proteomics *11,* 1177-1190.

Luevano, M., Bernard, H.U., Barrera-Saldana, H.A., Trevino, V., Garcia-Carranca, A., Villa, L.L., Monk, B.J., Tan, X., Davies, D.H., Felgner, P.L., and Kalantari, M. (2010). High-throughput profiling of the humoral immune responses against thirteen human papillomavirus types by proteome microarrays. Virology *405,* 31-40.

MacBeath, G., and Schreiber, S.L. (2000). Printing proteins as microarrays for high-throughput function determination. Science *289,* 1760-1763.

Margarit, I., Bonacci, S., Pietrocola, G., Rindi, S., Ghezzo, C., Bombaci, M., Nardi-Dei, V., Grifantini, R., Speziale, P., and Grandi, G. (2009). Capturing host-pathogen interactions by protein microarrays: identification of novel streptococcal proteins binding to human fibronectin, fibrinogen, and C4BP. Faseb j. *23,* 3100-3112.

Merbl, Y., and Kirschner, M.W. (2009). Large-scale detection of ubiquitination substrates using cell extracts and protein microarrays. Proc. Natl. Acad. Sci. U. S. A. *106,* 2543-2548.

Miller, J.C., Zhou, H., Kwekel, J., Cavallo, R., Burke, J., Butler, E.B., Teh, B.S., and Haab, B.B. (2003). Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. Proteomics *3,* 56-63.

Mueller, C., Liotta, L.A., and Espina, V. (2010). Reverse phase protein microarrays advance to use in clinical trials. Mol. Oncol. *4,* 461-481.

Nikitin, P.A., and Luftig, M.A. (2012). The DNA damage response in viral-induced cellular transformation. Br. J. Cancer *106,* 429-435.

Oh, Y.H., Hong, M.Y., Jin, Z., Lee, T., Han, M.K., Park, S., and Kim, H.S. (2007). Chip-based analysis of SUMO (small ubiquitin-like modifier) conjugation to a target protein. Biosens. Bioelectron. *22,* 1260-1267.

Orenes-Pinero, E., Barderas, R., Rico, D., Casal, J.I., Gonzalez-Pisano, D., Navajo, J., Algaba, F., Piulats, J.M., and Sanchez-Carbayo, M. (2010). Serum and tissue profiling in bladder cancer combining protein and tissue arrays. J. Proteome Res. *9,* 164-173.

Parsons, D.W., Li, M., Zhang, X., Jones, S., Leary, R.J., Lin, J.C., Boca, S.M., Carter, H., Samayoa, J., Bettegowda, C.*, et al.* (2011). The genetic landscape of the childhood cancer medulloblastoma. Science *331,* 435-439.

Paweletz, C.P., Charboneau, L., Bichsel, V.E., Simone, N.L., Chen, T., Gillespie, J.W., Emmert-Buck, M.R., Roth, M.J., Petricoin III, E.F., and Liotta, L.A. (2001). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. Oncogene *20,* 1981-1989.

Petricoin, E.F.,3rd, Espina, V., Araujo, R.P., Midura, B., Yeung, C., Wan, X., Eichler, G.S., Johann, D.J.,Jr, Qualman, S., Tsokos, M.*, et al.* (2007). Phosphoprotein pathway mapping: Akt/mammalian target of rapamycin activation is negatively associated with childhood rhabdomyosarcoma survival. Cancer Res. *67,* 3431-3440.

Poetz, O., Schwenk, J.M., Kramer, S., Stoll, D., Templin, M.F., and Joos, T.O. (2005). Protein microarrays: catching the proteome. Mech. Ageing Dev. *126,* 161-170.

Popescu, S.C., Popescu, G.V., Bachan, S., Zhang, Z., Gerstein, M., Snyder, M., and Dinesh-Kumar, S.P. (2009). MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays. Genes Dev. *23,* 80-92.

Popescu, S.C., Popescu, G.V., Bachan, S., Zhang, Z., Seay, M., Gerstein, M., Snyder, M., and Dinesh-Kumar, S.P. (2007). Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. Proc. Natl. Acad. Sci. U. S. A. *104,* 4730-4735.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R.*, et al.* (2005). Global analysis of protein phosphorylation in yeast. Nature *438,* 679-684.

Qiu, J., and LaBaer, J. (2011). Nucleic acid programmable protein array a just-in-time multiplexed protein expression and purification platform. Methods Enzymol. *500,* 151-163.

Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., Rosen, B., Lau, A.Y., Walter, J.C., and LaBaer, J. (2004). Self-assembling protein microarrays. Science *305,* 86-90.

Rapicavoli, N.A., Poth, E.M., Zhu, H., and Blackshaw, S. (2011). The long noncoding RNA Six3OS acts in trans to regulate retinal development by modulating Six3 activity. Neural Dev. *6,* 32-8104-6-32.

Sahiner, B., Chan, H.P., Petrick, N., Wagner, R.F., and Hadjiiski, L. (2000). Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Med. Phys. *27,* 1509-1522.

Salamat-Miller, N., Fang, J., Seidel, C.W., Smalter, A.M., Assenov, Y., Albrecht, M., and Middaugh, C.R. (2006). A network-based analysis of polyanion-binding proteins utilizing yeast protein arrays. Mol. Cell. Proteomics *5,* 2263-2278.

Sanchez-Carbayo, M., Socci, N.D., Lozano, J.J., Haab, B.B., and Cordon-Cardo, C. (2006). Profiling bladder cancer using targeted antibody arrays. Am. J. Pathol. *168,* 93-103.

Schnack, C., Hengerer, B., and Gillardon, F. (2008). Identification of novel substrates for Cdk5 and new targets for Cdk5 inhibitors using high-density protein microarrays. Proteomics *8,* 1980-1986.

Schweitzer, B., Roberts, S., Grimwade, B., Shao, W., Wang, M., Fu, Q., Shu, Q., Laroche, I., Zhou, Z., Tchernev, V.T.*, et al.* (2002). Multiplexed protein profiling on microarrays by rolling-circle amplification. Nat. Biotechnol. *20,* 359-365.

Shafer, M.W., Mangold, L., Partin, A.W., and Haab, B.B. (2007). Antibody array profiling reveals serum TSP-1 as a marker to distinguish benign from malignant prostatic disease. Prostate *67,* 255-267.

Shamay, M., Liu, J., Li, R., Liao, G., Shen, L., Greenway, M., Hu, S., Zhu, J., Xie, Z., Ambinder, R.F.*, et al.* (2012). A protein array screen for Kaposi's sarcoma-associated herpesvirus LANA interactors links LANA to TIP60, PP2A activity, and telomere shortening. J. Virol. *86,* 5179-5191.

Siprashvili, Z., Webster, D.E., Kretz, M., Johnston, D., Rinn, J.L., Chang, H.Y., and Khavari, P.A. (2012). Identification of proteins binding coding and non-coding human RNAs using protein microarrays. BMC Genomics *13,* 633-2164-13-633.

Skates, S.J., Horick, N., Yu, Y., Xu, F.J., Berchuck, A., Havrilesky, L.J., de Bruijn, H.W., van der Zee, A.G., Woolas, R.P., Jacobs, I.J., Zhang, Z., and Bast, R.C.,Jr. (2004). Preoperative sensitivity and specificity for early-stage ovarian cancer when combining cancer antigen CA-125II, CA 15-3, CA 72-4, and macrophage colony-stimulating factor using mixtures of multivariate normal distributions. J. Clin. Oncol. *22,* 4059-4066.

Smyth, G.K., and Speed, T. (2003). Normalization of cDNA microarray data. Methods *31,* 265-273.

Song, Q., Liu, G., Hu, S., Zhang, Y., Tao, Y., Han, Y., Zeng, H., Huang, W., Li, F., Chen, P.*, et al.* (2010). Novel autoimmune hepatitis-specific autoantigens identified using protein microarray technology. J. Proteome Res. *9,* 30-39.

Stillman, B.A., and Tonkinson, J.L. (2000). FAST slides: a novel surface for microarrays. Biotechniques *29,* 630-635.

Sutandy, F.X., Qian, J., Chen, C.S., and Zhu, H. (2013). Overview of protein microarrays. Curr. Protoc. Protein Sci. *Chapter 27,* Unit27.1.

Tao, S.C., Li, Y., Zhou, J., Qian, J., Schnaar, R.L., Zhang, Y., Goldstein, I.J., Zhu, H., and Schneck, J.P. (2008). Lectin microarrays identify cell-specific and functionally significant cell surface glycan markers. Glycobiology *18,* 761-769.

Tarrant, M.K., Rho, H.S., Xie, Z., Jiang, Y.L., Gross, C., Culhane, J.C., Yan, G., Qian, J., Ichikawa, Y., Matsuoka, T.*, et al.* (2012). Regulation of CK2 by phosphorylation and O-GlcNAcylation revealed by semisynthesis. Nat. Chem. Biol. *8,* 262-269.

Thao, S., Chen, C.S., Zhu, H., and Escalante-Semerena, J.C. (2010). Nepsilon-lysine acetylation of a bacterial transcription factor inhibits Its DNA-binding activity. PLoS One *5,* e15123.

Uchiyama, N., Kuno, A., Koseki-Kuno, S., Ebe, Y., Horio, K., Yamada, M., and Hirabayashi, J. (2006). Development of a lectin microarray based on an evanescent-field fluorescence principle. Methods Enzymol. *415,* 341-351.

Unfricht, D.W., Colpitts, S.L., Fernandez, S.M., and Lynes, M.A. (2005). Grating-coupled surface plasmon resonance: a cell and protein microarray platform. Proteomics *5,* 4432-4442.

Uzoma, I., and Zhu, H. (2013). Interactome mapping: using protein microarray technology to reconstruct diverse protein networks. Genomics Proteomics Bioinformatics *11,* 18-28.

Vazquez-Martin, A., Colomer, R., and Menendez, J.A. (2007). Protein array technology to detect HER2 (erbB-2)-induced 'cytokine signature' in breast cancer. Eur. J. Cancer *43,* 1117-1124.

Vidal, M., Brachmann, R.K., Fattaey, A., Harlow, E., and Boeke, J.D. (1996). Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. Proc. Natl. Acad. Sci. U. S. A. *93,* 10315-10320.

Vigil, A., Chen, C., Jain, A., Nakajima-Sasaki, R., Jasinskas, A., Pablo, J., Hendrix, L.R., Samuel, J.E., and Felgner, P.L. (2011). Profiling the humoral immune response of acute and chronic Q fever by protein microarray. Mol. Cell. Proteomics *10,* M110.006304.

Wegner, G.J., Lee, H.J., Marriott, G., and Corn, R.M. (2003). Fabrication of histidine-tagged fusion protein arrays for surface plasmon resonance imaging studies of protein-protein and protein-DNA interactions. Anal. Chem. *75,* 4740-4746.

Woodard, C., Shamay, M., Liao, G., Zhu, J., Ng, A.N., Li, R., Newman, R., Rho, H.S., Hu, J., Wan, J.*, et al.* (2012). Phosphorylation of the chromatin binding domain of KSHV LANA. PLoS Pathog. *8,* e1002972.

Yan, H., Park, S.H., Finkelstein, G., Reif, J.H., and LaBean, T.H. (2003). DNA-templated self-assembly of protein arrays and highly conductive nanowires. Science *301,* 1882-1884.

Young, K.H. (1998). Yeast two-hybrid: so many interactions, (in) so little time.. Biol. Reprod. *58,* 302-311.

Zhang, X., Zhu, S., Deng, C., and Zhang, X. (2012). An aptamer based on-plate microarray for high-throughput insulin detection by MALDI-TOF MS. Chem. Commun. (Camb) *48,* 2689-2691.

Zhang, Z., and Chan, D.W. (2010). The road from discovery to clinical diagnostics: lessons learned from the first FDA-cleared in vitro diagnostic multivariate index assay of proteomic biomarkers. Cancer Epidemiol. Biomarkers Prev. *19,* 2995-2999.

Zhang, Z., Yu, Y., Xu, F., Berchuck, A., van Haaften-Day, C., Havrilesky, L.J., de Bruijn, H.W., van der Zee, A.G., Woolas, R.P., Jacobs, I.J.*, et al.* (2007). Combining multiple serum tumor markers improves detection of stage I epithelial ovarian cancer. Gynecol. Oncol. *107,* 526-531.

Zheng, D., Wan, J., Cho, Y.G., Wang, L., Chiou, C.J., Pai, S., Woodard, C., Zhu, J., Liao, G., Martinez-Maza, O.*, et al.* (2011). Comparison of humoral immune responses to Epstein-Barr virus and Kaposi's sarcoma-associated herpesvirus using a viral proteome microarray. J. Infect. Dis. *204,* 1683-1691.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T.*, et al.* (2001). Global analysis of protein activities using proteome chips. Science *293,* 2101-2105.

Zhu, H., Cox, E., and Qian, J. (2012). Functional protein microarray as molecular decathlete: a versatile player in clinical proteomics. Proteomics Clin. Appl. *6,* 548-562.

Zhu, H., Hu, S., Jona, G., Zhu, X., Kreiswirth, N., Willey, B.M., Mazzulli, T., Liu, G., Song, Q., Chen, P.*, et al.* (2006). Severe acute respiratory syndrome diagnostics using a coronavirus protein microarray. Proc. Natl. Acad. Sci. U. S. A. *103,* 4011-4016.

Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A., Klemic, K.G., Smith, D., Gerstein, M., Reed, M.A., and Snyder, M. (2000). Analysis of yeast protein kinases using protein chips. Nat. Genet. *26,* 283-289.

Zhu, J., Gopinath, K., Murali, A., Yi, G., Hayward, S.D., Zhu, H., and Kao, C. (2007). RNA-binding proteins that inhibit RNA virus infection. Proc. Natl. Acad. Sci. U. S. A. *104,* 3129-3134.

Zhu, J., Liao, G., Shan, L., Zhang, J., Chen, M.R., Hayward, G.S., Hayward, S.D., Desai, P., and Zhu, H. (2009). Protein array identification of substrates of the Epstein-Barr virus protein kinase BGLF4. J. Virol. *83,* 5219-5231.

Zhu, X., Gerstein, M., and Snyder, M. (2006). ProCAT: a data analysis approach for protein microarrays. Genome Biol. *7,* R110.

Chapter 2: Scaffolding Proteins in Phosphorylation-Mediated Signal Transduction

2.1 Introduction

Scaffolding proteins are defined as proteins that interact with at least two other proteins in a signaling pathway. As such, they play a crucial role in the regulation of cellular signal transduction. Protein phosphorylation and dephosphorylation is an important means of regulation and protein function that occurs in both prokaryotic and eukaryotic organisms. The phosphorylation of a protein may result in a conformational change in its structure leading to its activation or deactivation, and thus sequential and recursive phosphorylation of proteins allows the transmission of extracellular signals to intracellular targets. One well-known example is the RAS-ERK pathway, in which a small G-protein *RAS* activates MAP3K *RAF*, which then phosphorylates and activates MAP2K *MEK1* (*MAPKK1*). *MEK1* then phosphorylates and activates MAPK *ERK* (Shaw et al., 2009). Biological systems contain a large number of phosphorylation-related signaling pathways and proteins in different pathways may overlap with each other. This extensive cross-talk leads to the question of specificity, and understanding how each pathway is distinctly regulated is of great importance to the biological community. We believe the answer may partially lie in the existence of scaffolding proteins.

Scaffolding proteins act as a sort of "molecular glue," linking multiple components in a pathway to facilitate signal transduction, and as such play crucial roles in the regulation of signaling cascades. They can enhance signaling specificity by sequestering proteins, preventing cross-talk between proteins in different signaling pathways. They can also increase the signaling efficiency by increasing the local

concentration of each signaling component. Thus, the knowledge of scaffolding proteins can help improve our understanding of the regulation of cellular signal transduction (Levchenko et al, 2000).

2.2 Prediction of Scaffolding Proteins in Signaling Cascades

For a given protein pair, we calculated the distances for all possible paths connecting them in the protein-protein interaction (PPI) network. A distance of 1 indicates that two proteins directly interact with each other, while a distance of 2 indicates that they both interact with a third protein. Among 1,103 protein pairs with known kinase-substrate relationships (KSRs), 24.9% of them have a distance of 2 in the PPI network. In contrast, of the 6.4×10^7 human protein pairs in the PPI network, only 2.7% have a distance of 2 (Figure 2.1). This stark enrichment suggests that protein mediators may play an important role in kinase signaling cascades.

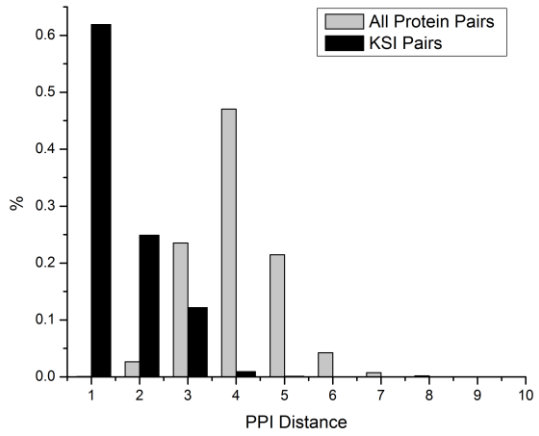

Figure 2.1 KSI pairs are significantly enriched in PPI distance=2. 24.9% of KSI pairs have PPI distance of 2, while only 2.7% of PPI pairs have PPI distance of 2. The PPI distance of a protein pair is defined as the shortest distance of the two proteins in PPI network.

In order to predict potential scaffolding proteins in phosphorylation signaling cascades, the PPI and KSR networks were superimposed (Figure 2.2). This was based on the assumption that a scaffolding protein interacts with most components in a phosphor-signaling pathway.

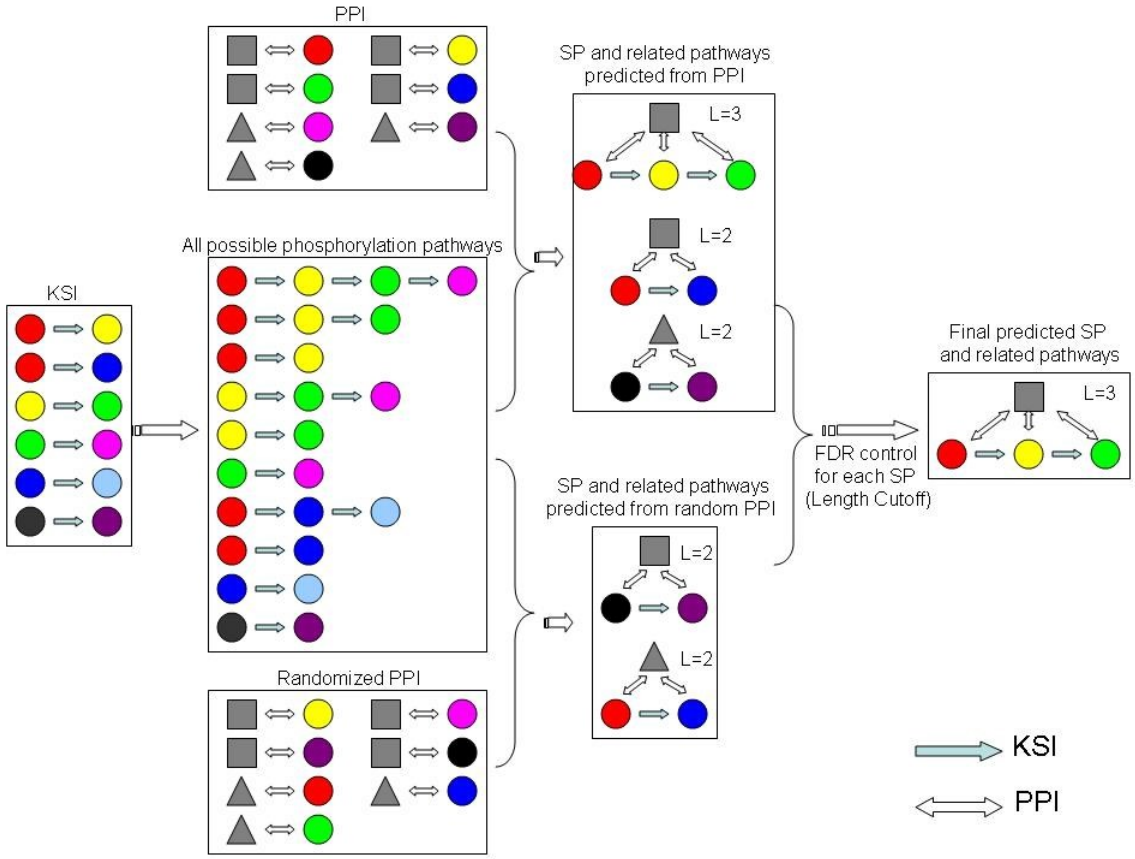

Figure 2.2 Schematic diagram of procedure to predict scaffolding proteins and related pathways. The randomized PPI is produced from real PPI by randomly selecting two PPI pairs and change their partners. The list of all possible phosphorylation pathways are produced by DFS searching algorithm. See method for details.

Here, the pathway is defined as a set of proteins with linear KSRs. For example, if protein A phosphorylates protein B, and protein B phosphorylates protein C, we constructed a pathway of A → B → C. Continuous sub-paths within a long pathway will also be considered as separate pathways (such as A → B and B → C). Note that such defined pathways are not necessary to be the same biological pathway as those defined in other databases (eg. KEGG database).

A very stringent requirement was made in predicting potential scaffolding proteins by examining whether any protein interacts with *all* components in a given pathway. Proteins with many known PPIs (eg. hub proteins) have a high chance of being scaffolding proteins. In addition, short pathways tend to share common PPI partners. To correct for these bias, simulations were performed to estimate the statistical significance for the prediction of scaffolding proteins. Random PPI networks were generated by permuting the PPIs, while keeping the interaction degree (i.e. number of interacting partners) for each protein. In the random PPI networks, we calculated the chance that a protein predicted as a scaffolding protein for a pathway with the interaction degree of the protein and the length of the pathway as variables.

Choosing 0.01 as a False Discovery Rate cutoff, 212 proteins were predicted as scaffolding proteins, which target to 612 non-redundant phosphorylation pathways. Among the 1,103 known KSRs, 359 of them (33%) are associated with at least one predicted scaffolding protein. The resulting network is shown in Figure 2.3.

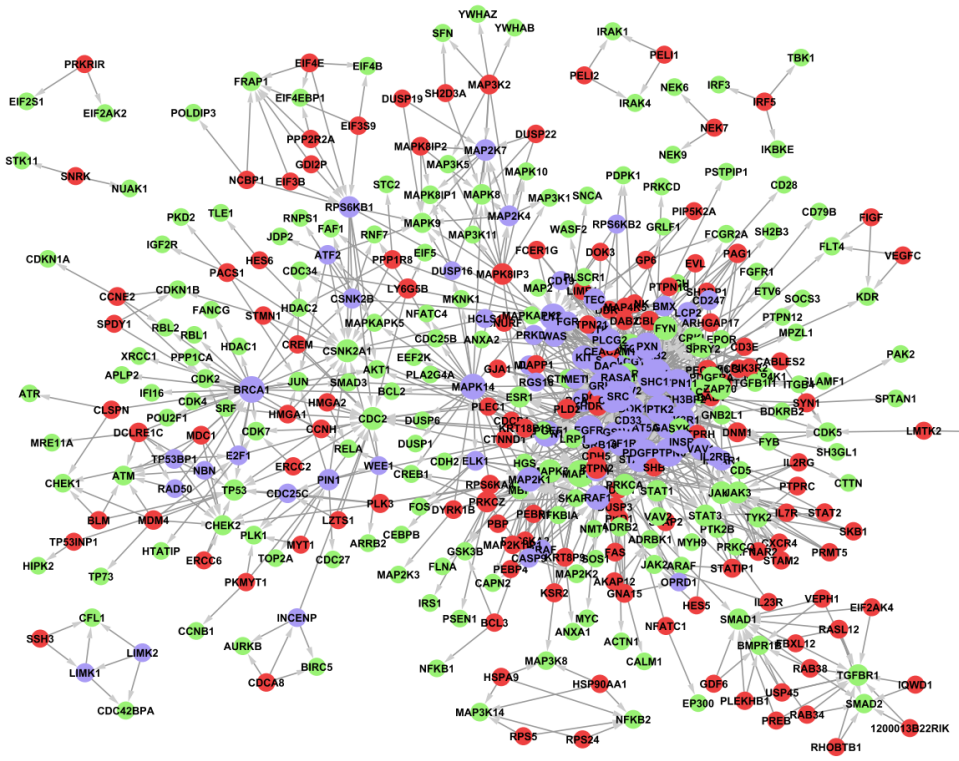

Figure 2.3. Network of predicted PSPs and proteins in their associated pathways. Scaffolding proteins were colored as red and proteins in pathways were colored as green. Scaffolding proteins were colored as blue if they were also proteins in pathways. Edge arrows were from scaffolding proteins to proteins in its related pathways. The size of node is proportional to its degree in network.

We then examined whether these scaffolding proteins are simply due to the high interaction degrees (numbers of interacting neighbors in the PPI network). Based on the PPI degree distribution, we found that the peak of the distribution locates around 10; on average, each scaffolding protein interacts with 10 proteins in PPI networks (Figure 2.4 A). This distribution is similar to that of known scaffolding proteins (not shown). This result indicates the prediction of scaffolding proteins is unlikely to be an artifact due to their high PPI degrees, whereas, we did observed that proteins with high PPI degrees have high possibilities to be scaffolding proteins.

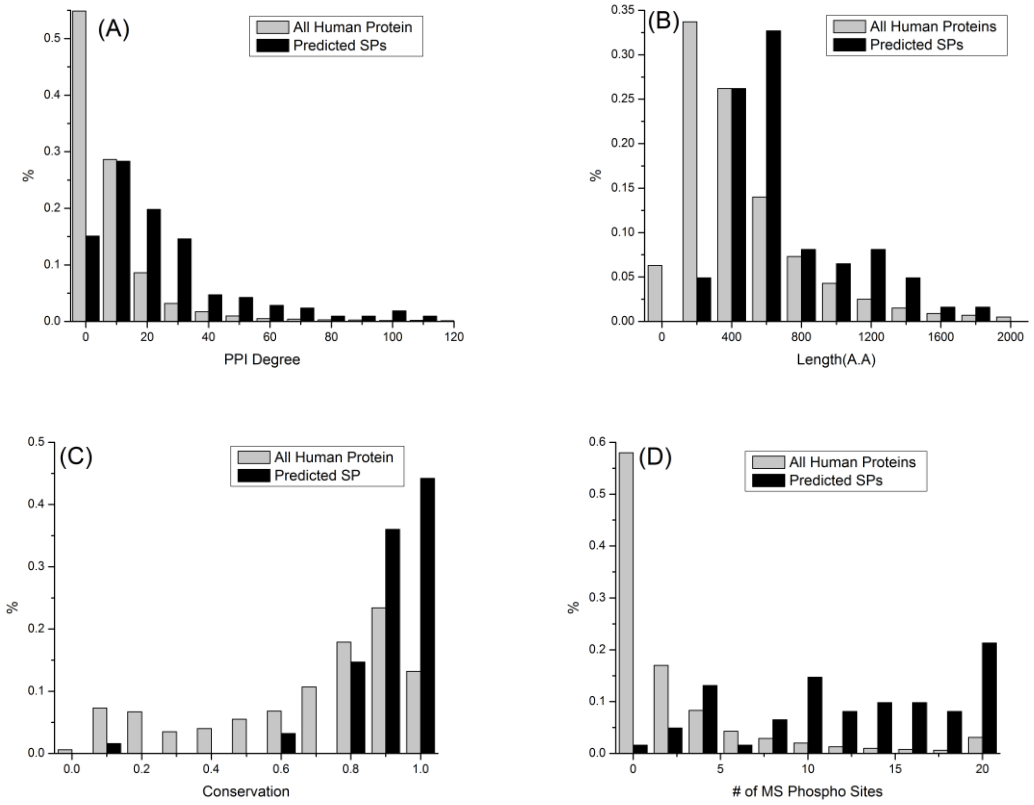

Figure 2.4 Statistics of predicted scaffolding proteins (PSPs). As a comparison, the distribution of background (all human proteins) was also drawn. (A) Distribution of PPI degree. (B) Distribution of protein length. (C) Distribution of protein conservation. The conservation was computed from the comparison with mouse proteome by BlastP. (D) Distribution of MS Phospho site numbers.

We observed that the occurrences of the scaffolding network motifs (i.e. one protein interacting with all components in a pathway) are significantly enriched compared to their occurrences in the networks where the PPIs were randomly permutated. For example, the scaffolding network motif with a phosphorylation cascade length of 6 occurs 18 times, whereas we expect in only 0.26 times in a randomized network. This suggests that a scaffolding mediator is a widely-used mechanism in kinase signaling cascades.

We collected 78 possible scaffolding proteins for kinase signaling pathways through literature curation. Our prediction recovered 18 of them, which is more than a 3-fold enrichment ($p<1x10^{-5}$, hypergeometric distribution).

Of the 612 scaffolding mediated phosphorylation pathways, 414 (68%) are associated with only one scaffolding protein, suggesting that pathways are often specifically regulated by only one scaffolding protein (Figure 2.5 A). Instead, 61% of scaffolding proteins are associated with more than one pathway, suggesting that scaffolding proteins can participate in multiple pathways (Figure 2.5 B).

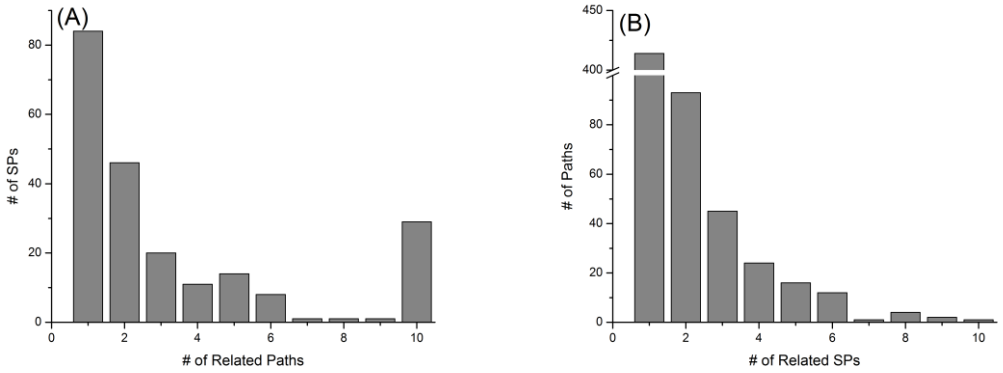

Figure 2.5. Specificity of PSPs and pathways. (A) Number of PSPs related to paths. 31.3% of PSPs relates to only one path, but there are 37 PSPs related to no less than 20 paths. (B) Number of pathways related to PSPs. 3229 paths (83.9%) related to only one PSP.

Some partially overlapped pathways are involved in different biological processes and can be regulated by different scaffolding proteins. For example, the MAPK signaling pathway *RAF1* → *MAP2K1* → *MAPK1* is regulated by the scaffolding protein *KSR2*. The MAPK pathway is partially overlapped with the T-cell receptor signaling pathway, *SRC* → *RAF1* → *MAP2K1*, which is regulated by scaffolding protein *MAPK8IP3*. This example demonstrates that scaffolding proteins can provide specificity to the signaling pathways and prevent possible undesired crosstalk between pathways.

2.3 Characterization of Scaffolding Proteins

We first examined the gene ontology (GO) annotation associated with the predicted scaffolding proteins. The GO biological process analysis indicates that 106 of the 212 predicted scaffolding proteins are associated with the GO term "signal transduction" ($p<1x10^{-28}$, hypergeometric distribution), and 75 of them are annotated to be related to "intracellular signaling cascade" ($p<1x10^{-32}$, hypergeometric distribution), both over three times enriched than expected (Figure 2.6). Furthermore, 38 of predicted scaffolding proteins are associated with the GO term "regulation of phosphorylation," and 36 with "protein kinase cascade." Although our prediction did not use these GO terms as an input, the GO analysis of the potential scaffolding proteins suggests the validity of our prediction.

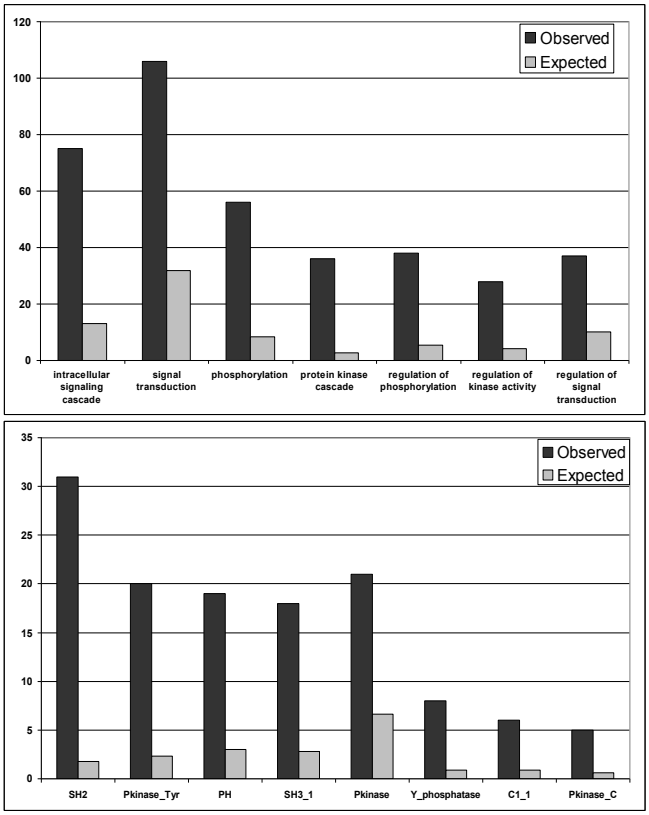

Figure 2.6. GO and Pfam analysis of PSPs. The GO and Pfam terms are sorted increasingly from left to right by p-value.

To characterize the predicted scaffolding proteins, we also examined the protein domains in these proteins as defined in Pfam. Compared to the expected occurrence of the corresponding domains, we found the predicted scaffolding proteins are enriched for several protein domains (Figure 2.6). Many enriched domains are known to interact with phosphorylation sites and play a role in signaling cascades, including SH2, SH3, 14-3-3, PH and Ank. Interestingly, kinase domains are also enriched, such as Pkinase and Pkinase_Tyr, suggesting that some scaffolding proteins are kinases themselves. As a side note, the two scaffolds used in this work (*PIN1* and *ATF2*) did not have kinase domains enriched. Also of note, scaffolding proteins are generally large proteins because they need to interact with different proteins. The comparison between predicted scaffolding proteins and the human proteome shows the predicted scaffolding proteins are significantly larger than that of background (average 600 residues for scaffolding proteins vs. 200 residues for all human proteins) (Figure 2.4 B). However, some known and predicted PSPs are small proteins indeed, as they can form large complexes of polymer, such as ISCU (Chandramouli et al, 2007).

If scaffolding proteins are essential for signaling pathways, we expect that these proteins are under evolutionary constraint as well. By comparing with mouse protein sequences, we calculated the conservation score for each protein in humans. The predicted scaffolding proteins have an average conservation score of 0.90, while the average conservation score for all human proteins is 0.68 (Figure 2.4 C). In fact, 92% of predicted scaffolding proteins have conservation scores larger than 0.8, while on average we only expect that 47% of human proteins have that level of conservation.

In an attempt to understand the interaction mode between scaffolding proteins and the pathways they mediate, we studied the phosphorylation sites on the predicted scaffolding proteins. We collected 31,000 known phosphorylation sites obtained from mass spectrometry experiments and then mapped these sites on the proteins. We found a majority of predicted scaffolding proteins (98%) contain at least one known phosphorylation site. In contrast, only 42% of proteins in the entire human proteome contain known phosphorylation sites (Figure 2.4 D). The high enrichment in phopho-proteins implies that the interactions between scaffolding proteins and proteins in signaling cascades might be phosphorylation dependent.

2.4 Validation of Scaffolding Proteins Using Protein Microarrays Shows ATF2 and PIN1 Mediated Phosphorylation

In order to validate the predicted scaffolding proteins, kinase assays were carried out on the human proteome array, which contains over 17,000 full-length human proteins (Jeong et al., 2012). Predicted scaffolding proteins were selected based on the ability to purify kinases with good activity, and contained kinase pathways that were easily inducible/inhibited for future *in vivo* validations (Table 2.1).

Predicted Scaffold	**Kinase**
PIN1	CSNK2A1
ATF2	CSNK2A1
	MAPK9
CBL*	Src
GRB2*	Src

Table 2.1 Validation hit list of predicted scaffolding proteins. *Known scaffolding protein.

Kinases and scaffolds were purified from yeast as GST-fusion proteins and shown to have high quality by coomassie staining (Figure 2.7). Dot blot assays were performed

to assess kinase activity using a mix of commonly phosphorylated proteins as a substrate (*Histone H2, Histone H3, Casein,* and *MBP*) in the presence of ^{32}P-γ-ATP. Reactions were carried out for 30 minutes at 30°C before spotting onto nitrocellulose. Commercial PKA (1U/μL) was used as a positive control and negative control contained no enzyme. After drying, the nitrocellulose was washed 3 times with PBS and dried again. The membrane was exposed overnight to film before developing. All three kinases showed what we consider to be very good activity. Arrays were treated in duplicate with kinases with and without scaffold in the presence of ^{33}P-γ-ATP for 30 minutes at 30°C in a humidity chamber. After incubation, the arrays were quickly immersed in two 400 mL beakers filled with TBST before washing 3 times with TBST for 10 minutes followed by 3 washes with 0.5% SDS for 10 minutes. Arrays were quickly immersed in preheated 37°C ddH_2O before being spun dry by centrifugation (2 minutes at 2000 rpm). Microarrays were placed in a cassette and exposed to film for 30 days at -80°C.

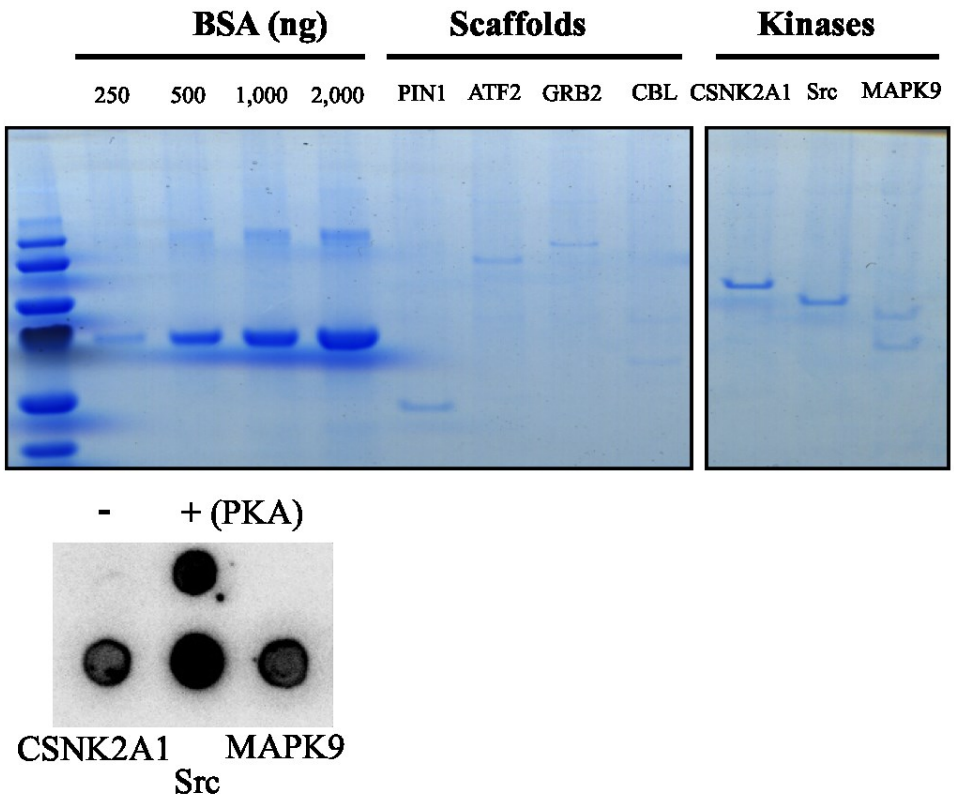

Figure 2.7 Coomassie stain of scaffolding proteins and kinases shows high quality of purification. Dot blot using generic substrate mix indicates that kinases are very active against generic substrates.

After 30 days, the film was developed and scanned before careful analysis using GenePix software. To qualify as a true hit, stringent criteria were employed. Each array was carefully aligned and each hit was identified manually by eye. True hits had a F/B ratio greater than 1.5, were present in both replicates, and not present in the control arrays. After this analysis, 28 unique hits were discovered for the *CSNK2A1* and *MAPK9* experiments (Table 2.2). To validate these hits, each spot was manually examined by eye and shown to be consistent with our conclusions (Figure 2.8). All in all, this marks a promising start in identifying scaffold-mediated protein phosphorylation and further *in vivo* studies will determine the scope of our observations.

CSNK2A1 Alone	CSNK2A1 + ATF2	CSNK2A1 + PIN1	MAPK9 Alone	MAPK9 + ATF2
N=2	N=4	N=17	N=2	N=9
PDCD4 IOH11605	ZNF554 IOH26717	BC036974 BC036974	PPT2 IOH3929	FLJ22639 IOH12793
CENPB IOH29042	**BNIPL1 IOH11810**	SIRT1 IOH11855	SIRT4 IOH59335	KIR2DL*4* IOH28039
	C2orf13 IOH26160	C3orf37 IOH13046		CENPB IOH29042
	WASF2 IOH26184	CSNK2A1 IOH13704		KLK14 IOH35361
		MCC IOH14306		PSME1 IOH3647
		C19orf33 IOH14605		HELLS IOH44165
		RP6-213H19.1 IOH39650		MRPL18 IOH4500
		C13orf15 IOH40124		CASP4 IOH45758
		CDC123 IOH4905		IGL@ IOH62887
		TBC1D2 IOH63057		
		SLC4A1AP IOH63362		
		N.D. IOH6539		
		KIAA1143 IOH7405		
		SSX3 IOH7517		
		BNIPL1 IOH11810		
		C2orf13 IOH26160		
		WASF2 IOH26184		

Table 2.2 List of *PIN1* and *ATF2* meditated phosphorylation events. Each array was manually aligned and scored Final hits that were identified were replicated, not present in control arrays, and had F/B ratios >1.5. Proteins in bold were hit with both *ATF2* and *PIN1*.

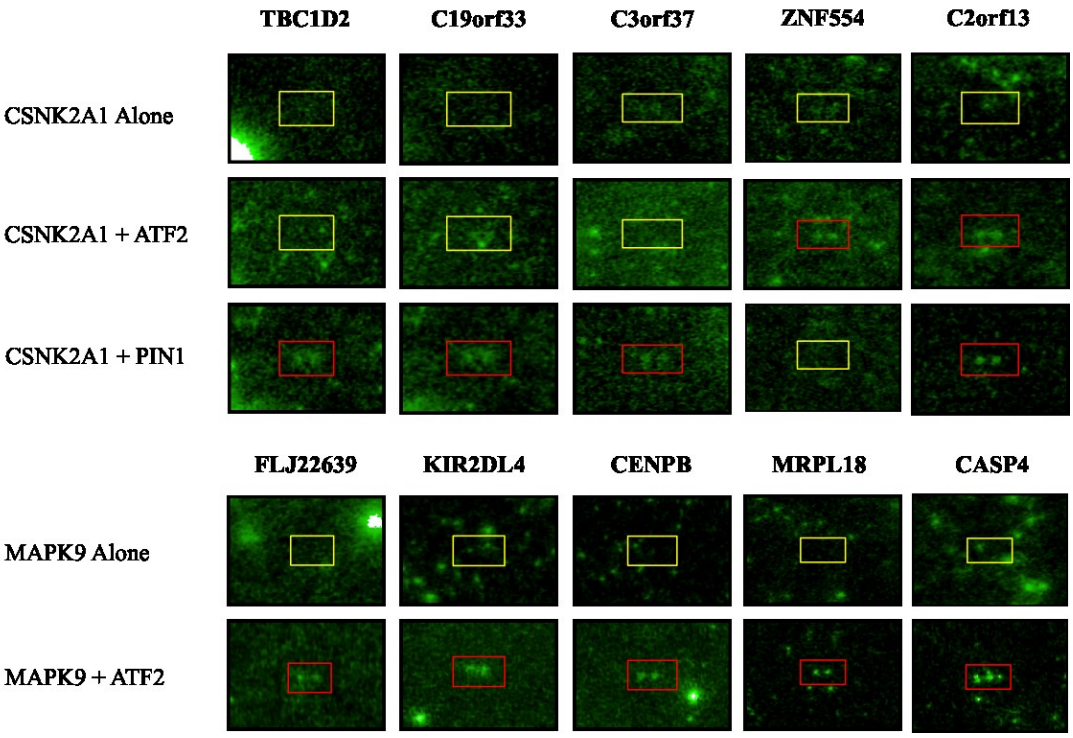

Figure 2.8 Microarray validations for *CSNK2A1* and *MAPK9*. Red borders indicate positive hit.

2.5 Discussion

Signal transduction by phosphorylation is the most universal and well-studied mechanism that cells employ to mediate cell signaling, although there are still many interactions to discover. Kinase substrates have long been identified from their consensus sequence motifs, an amino acid sequence uniquely recognized by a particular kinase in which phosphorylation occurs. However, with the accumulation of experimental data, many kinases break this long-held rule and have been found to share similar phosphorylation motifs with other kinases, but identify totally different sets of substrates. This phenomenon has plagued many paradigms and can only be partially explained by differences in sub-cellular localization of kinase and substrate. Another recent explanation is in the form of protein docking and scaffolding proteins. Docking has frequently been used to indicate the binding of small molecule drug candidates to their protein targets, but can also be employed to indicate the weak and transient interactions between proteins (Halperin et al, 2002). Recent publications indicate that docking plays an essential role in the recognition of the substrate by a kinase. For example, *GSK3* kinase-catalyzed protein phosphorylation often requires a stable kinase-substrate docking interaction, and kinase *BIN2* directly interacts with a 12 amino acid motif adjacent to the C-terminus of its substrate *BZR1* in order to recognize it (Peng et al, 2010). Currently only a few docking motifs are known for kinases, but accumulation of these kinase docking motifs are critical in developing accurate prediction methods.

This work is a first trial attempt at large-scale projections of predicted scaffolding proteins (PSPs). We have developed a method to predict 212 PSPs we believe are of high

confidence. Relative to other proteins, PSPs are generally larger and have conserved protein sequences. They also contain more phosphorylation sites, indicating their scaffolding function may be phosphorylation dependent. Although proteins with high PPI degrees are more likely to be PSPs, not all proteins with high PPI degrees are predicted as such, indicating the predictions are less likely to be random. Initial microarray validations have provided a first start in validating this method. Twenty-eight substrates were found to be phosphorylated only in the presence of a predicted scaffold. Further studies will be necessary to further characterize these initial observations. The predicted PSPs are dependent on the ever expanding KSI and PPI databases, which, as they continue to accumulate more data, will be able to predict more PSPs in the future.

2.6 Materials and Methods

Protein-protein Interactions

Human protein-protein interaction (PPI) data was collected from three databases: IntAct (ftp://ftp.ebi.ac.uk/pub/databases/intact/current/), HPRD (HPRD_Release_7_09012007, http://www.hprd.org/sentDataRequest) and BioGRID (biogrid-all-2.0.45.tab, http://www.thebiogrid.org/downloads.php). These data were then formatted and reorganized to remove redundancies. In total, we obtained 55,048 human PPI pairs.

Computation of PPI Distance

The PPI distances of a protein pair is defined as the shortest distance of the protein pair in the PPI network, and can be computed using Breadth-First Search (BFS) algorithm. We took each protein with PPI information as a root, and defined it as the first level of a tree. We then extended the root to take all its neighbors as the nodes at the second level of the tree. We next collected and took the neighbors of all nodes at second level as the nodes at the third level of the tree, and all nodes that had appeared in previous levels would be deleted. We repeated this procedure till no further level could be added to the tree. This resulted in the PPI distances between root node and all other nodes in the tree being the difference of their levels. For example, the PPI distance between root node (first level) and a node at the fourth level is 3. This allowed us to obtain the minimal distance of each protein pair.

Identification of Phosphorylation-related PSPs

We took each kinase as root, and extended its substrates one by one by a Depth-First Search (DFS) algorithm. Each path starting from the root in the tree represents a possible phosphorylation pathway or its continuous substring. Here, we require the path must start from a root node, but does not need to end at leaf node. The minimum length of a pathway was set as 2. By doing it this way, we can list all possible phosphorylation pathways and their continuous substrings and remove any redundancies. We also included the continuous substrings of long pathways because they may not have corresponding PSPs, while its continuous substrings do.

Protein Microarray Assay

Proteins were purified, printed, and analyzed as described previously (Jeong et al, 2012). Kinases and scaffolds ORFs were expressed as GST-fusion proteins in yeast. Cultures (50 mL) were grown at 30° C to OD_{600} 1.0-1.2 and induced with 2% galactose for 4-6 hours. Harvested cells were lysed with glass beads in lysis buffer (100 mM Tris-HCl [pH 7.4], 100 mM NaCl, 1 mM EGTA, 0.1% 2-mercaptoethanol, 0.5 mM PMSF, 0.1% Triton X-100, protease inhibitor cocktail [Roche], and phosphatase inhibitor cocktails 2 and 3 [Sigma]). GST-proteins were bound to glutathione beads (GE healthcare) for 40 minutes at 4° C and washed 3 times with Wash Buffer I (50 mM Tris-HCl [pH 7.4], 500 mM NaCl, 1 mM EGTA, 10% glycerol, 0.1% Triton X-100, 0.1% 2-mercaptoethanol, and 0.5 mM PMSF) and 3 times with Wash Buffer II (50 mM HEPES [pH 7.4], 100 mM NaCl, 1 mM EGTA, 10% glycerol, 0.1% 2-mercaptoethanol, and 0.5 mM PMSF) before 2 30 minute elutions in elution buffer (100 mM Tris-HCl [pH 8.0], 100 mM NaCl, 10 mM $MgCl_2$, 30 mM glutathione, and 20% glycerol). Eluate was collected and concentrations were determined through BSA standard.

Protein chips were briefly dipped in TBS to remove excess glycerol from printing procedure before blocking in 3 mL of blocking buffer (3% BSA in TBST) for 1 hour. Chips were washed 3 times in TBST before the addition of 125 uL of kinase buffer containing 3:1 scaffold:kinase in kinase buffer (50 mM Tris-HCl [pH 7.5], 100 mM NaCl, 10mM $MgCl_2$, 1 mM $MnCl_2$, 1 mM DTT, 1 mM EGTA, 25 mM HEPES-KOH [pH 7.5], 1 mM $NaVO_4$, 1 mM NaF, 0.1% NP-40, 0.0000556 mM ^{33}P-γ-ATP [Perkin Elmer; 2 μL/array]). Chips were placed in a humidity chamber and incubated for 30 minutes at 30° C. Following the reaction, chips were quickly immersed in two separate

beakers of TBST and washed 3 times in TBST for 10 minutes followed by 3 washes in 0.5% SDS for 10 minutes. Arrays were then quickly dunked in water heated to 37° C and dried by centrifugation before being arranged in a standard film cassette and exposed to film (Kodak BioMax MR) for 30 days. Cassette was stored at -80° C.

Dot Blot Assay

2 μL purified kinases were mixed with 1 μL substrate mix (1:1:1 casein:MBP:Histone H3 100 ng/μL dissolved in TBS) and 2 μL 2.5x reaction buffer (90 mM Tris-HCl, pH 7.5, 180 mM NaCl, 9 mM $MgCl_2$, 0.9 mM $MnCl_2$, 0.9 mM DTT, 9μM cold ATP, 2.5 mM EGTA, 20 mM HEPES-KOH, pH 7.5, 0.9 mM NaF, 0.9 mM Na_3VO_4, and 5.954E-05 mM ^{32}P-γ-ATP [Perkin Elmer; 0.2 μL/5.6 μL reaction mix]) and incubated at 30° C for 30 minutes. Reactions were quenched by spotting entire mix onto nitrocellulose paper and drying for 15 minutes. Membrane was then washed 3 times for 10 minutes with PBS and dried again for 15 minutes. Blots were exposed to film overnight.

2.7 References

Bhattacharyya, R.P., Remenyi, A., Yeh, B.J., Lim, W.A. (2006) Domains, motifs, and scaffolds the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem *75*, 655-80.

Chandramouli, K., et al. (2007). Formation and properties of [4Fe-4S] Clusters on the IscU scaffold protein. Biochemistry *46*, 6804-11.

Halperin, I., et al. (2002). Principles of Docking: An overview of search algorithms and a guide to scoring functions. Proteins *47,* 409-443.

Jeong, J.S., Jiang, L., Albino, E., Marrero, J., Rho, H.S., Hu, J., Hu, S., Vera, C., Bayron-Poueymiroy, D., Rivera-Pacheco, Z.A.*, et al.* (2012). Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. Mol. Cell. Proteomics *11,* O111.016253.

Levchenko, A., Bruck, J., Sternberg, P.W. (2000). Scaffold proteins may biphasically affect the level of mitogen-activated protein kinase signaling and reduce its threshold properties. PNAS *97*, 5818-23.

Newman RH, Hu J, Rho H-S, Xie Z, Woodard C, Neiswinger J, Hwang W, Shirley M, Hu S, Cooper C, Jeong JS, Wu G, Lin J, Gao X, Ni Q, Dalby K, Ji H, Desiderio S, Birnbaum MJ, Cole PA, Knapp S, Ryazanov A, Zack DJ, Blackshaw S, Pawson T, Gingras A-C, Pandey A, Turk BE, Zhang J, Zhu H, Qian J (2013). Construction of Human Activity-Based Phosphorylation Networks. Molecular Systems Biology *9,* 655

Peng Peng, et al. (2010). A direct docking mechanism for a plant GSK3-like kinase to phosphorylate its substrates. JBC *285*, 24646-53.

Shaw, A.S., Filbert, E.L. (2009). Scaffold proteins and immune-cell signaling. Nature Reviews Immunology *9*, 47-56.

Chapter 3: Profiling the O-GlcNAcylation of the Kinome Using Protein Microarray Technology

3.1 Introduction

Signal transduction in cells largely relies on the large, dynamic mix of inducible post-translational modifications (PTMs) of proteins. Whether phosphorylation, ubiquitylation, acetylation, or nitrosylation, the list of modifications and substrates tied to them continues to grow. While phosphorylation has certainly been the most well-studied PTM, especially as it relates to serine and threonine modification, recently proteins modified by β-N-acetylglucosamine (*O*-GlcNAc) have been emerging as prominent players in regulating cell signaling, cell division, transcription, and metabolism (Hart et al., 2007; Zeidan et al., 2010). The cycling of *O*-GlcNAc on proteins is facilitated by two enzymes, *O*-GlcNAc transferase (OGT), which catalyzes the addition of *O*-GlcNAc, and *O*-GlcNAcase (OGA), which removes *O*-GlcNAc through hydrolysis. The dynamics between these two enzymes is regulated through various stimuli, and the myriad of proteins that have been detected with this modification is just as varied. Transcription factors, histones, cytoskeletal proteins, and some kinases have all been identified as being *O*-GlcNAcylated (Hart et al., 2011). The effect of *O*-GlcNAcylation on proteins is varied as well. One of the more well-characterized examples is the modulation of calcium/calmodulin-dependent kinase IV (*CaMKIV*) activity by it's *O*-GlcNAcylation (Dias et al., 2009; Song et al., 2008; Slawson et al., 2010). In its basal/inactive state, *CaMKIV* is heavily glycosylated. Upon stimulation and calcium influx, S189 *O*-GlcNAcylation is removed by OGA, exposing the active loop site, allowing *CaMKK* to phosphorylate *CaMKIV* at S200 and activate it. Oftentimes, *O*-GlcNAcylation sites are

also phosphorylation sites, as is the case with nitric-oxide synthase (Du et al., 2001), c-Myc (Kamemura et al., 2002), and RNA polymerase II (Comer et al., 2001). However, this reciprocal relationship is not always seen. Several proteins can be both *O*-GlcNAcylated and phosphorylated at distant sites, or modification on adjacent sites can alter the modification of the other (Yang et al., 2006).

Recently, studies have altered the assumption that *O*-GlcNAc and phosphorylation are simply antagonistic toward each other by exhibiting that while increasing global *O*-GlcNAcylation did indeed decrease a large number of phosphorylation events, it also *increased* phosphorylation at many other sites (Wang et al., 2008). This implies that although *O*-GlcNAcylation may compete with kinases for site modification, it may also modify kinases and phosphatases as well. This leads to a great void in understanding how OGT/OGA and kinases/phosphatases are regulated and a high throughput approach to identify kinase substrates of OGT would be very useful in understanding the reach of *O*-GlcNAcylation on the kinome. Initial studies in the Hart lab used [H^3] radiolabeling on a small functional microarray identified 42 kinases as being *O*-GlcNAcylated (Dias et al., 2012). However, the severe sensitivity limitations of the assay left something to be desired. This led to a collaboration between the Zhu and Hart labs and the development of a three-pronged approach to probe the kinome using protein microarray technology.

3.2 Kinome Collection, Purification and Microarray Fabrication

In order to study the *O*-GlcNAcylation of kinases, a kinome collection was consolidated from the Invitrogen ORF collection. After Pfam analysis to eliminate any clones that did not possess a canonical kinase domain, a final list was obtained that contained 350 unique kinases, representing about 70% of the human kinome (Table 3.1). The collection contains kinases from almost all kinase families, including AGC, CAMK, CK1, CMGC, STE, TK, and TKL. Each kinase was expressed in yeast as a GST-fusion protein and affinity purified using glutathione sepharose beads (GE Healthcare). Protein purity and amounts were assessed using Coomassie staining against BSA standards (Figure 3.1).

Clone ID	***Gene Symbol***	*Clone ID*	***Gene Symbol***	*Clone ID*	***Gene Symbol***	*Clone ID*	***Gene Symbol***	*Clone ID*	***Gene Symbol***
IOH5262	***AAK1***	*IOH11766*	***CSNK1A1L***	*IOH29960*	***MAP2K1***	*IOH26671*	***PAK1***	*IOH36120*	***RPS6KA6***
IOH40442	***ABL2***	*IOH4975*	***CSNK1D***	*IOH3639*	***MAP2K2***	*IOH2475*	***PAK4***	*IOH29907*	***RPS6KB1***
IOH21860	***ACVR1***	*IOH21160*	***CSNK1E***	*IOH21715*	***MAP2K3***	*IOH6311*	***PAK4***	*IOH29003*	***RPS6KB1***
IOH4418	***ACVR1B***	*IOH21026*	***CSNK1G1***	*IOH29276*	***MAP2K4***	*IOH21457*	***PAK4***	*IOH6536*	***RPS6KB2***
IOH10813	***ACVR1C***	*IOH10417*	***CSNK1G2***	*IOH28657*	***MAP2K4***	*IOH20961*	***PAK6***	*IOH4738*	***RPS6KB2***
IOH29719	***ACVR2A***	*IOH27760*	***CSNK1G2***	*IOH6702*	***MAP2K5***	*IOH12390*	***PBK***	*IOH6324*	***RPS6KL1***
IOH63235	***ACVR2B***	*IOH26429*	***CSNK1G3***	*IOH59078*	***MAP2K6***	*IOH6258*	***PCTK1***	*IOH63009*	***SCYL1***
IOH63228	***ACVR2B***	*IOH13704*	***CSNK2A1***	*IOH27169*	***MAP2K7***	*IOH4605*	***PCTK1***	*IOH61176*	***SCYL2***
IOH21048	***ACVRL1***	*IOH6369*	***CSNK2A2***	*IOH9668*	***MAP3K11***	*IOH9712*	***PCTK3***	*IOH13493*	***SCYL3***

IOH29100	***ADCK1***	*IOH21981*	***DCAMKL2***	*IOH45191*	***MAP3K13***	*IOH27481*	***PCTK3***	*IOH3192*	***SGK***
IOH22636	***ADCK5***	*IOH20970*	***DCLK1***	*IOH27594*	***MAP3K14***	*IOH22618*	***PDGFRB***	*IOH14023*	***SGK2***
IOH25981	***ADRBK1***	*IOH5763*	***DDR1***	*IOH21077*	***MAP3K7***	*IOH12032*	***PDIK1L***	*IOH35747*	***SGK2***
IOH56275	***ADRBK2***	*IOH28909*	***DDR2***	*IOH29760*	***MAP3K7***	*IOH10145*	***PDK3***	*IOH41642*	***SGK3***
IOH3692	***AKT1***	*IOH40696*	***DMPK***	*IOH37873*	***MAP3K7***	*IOH25726*	***PDK4***	*IOH29894*	***SNF1LK***
n/a	***AKT2***	*IOH14793*	***DYRK1B***	*IOH29592*	***MAP3K8***	*IOH14589*	***PDPK1***	*IOH45349*	***SNF1LK2***
IOH43070	***AKT3***	*IOH2412*	***DYRK2***	*IOH26933*	***MAP4K2***	*IOH27406*	***PDPK1***	*IOH61637*	***SNRK***
IOH37801	***AKT3***	*IOH6475*	***DYRK2***	*IOH45342*	***MAP4K2***	*IOH37739*	***PFTK1***	*IOH12563*	***SRC***
IOH7002	***ALS2CR2***	*IOH61505*	***DYRK3***	*IOH27202*	***MAP4K5***	*IOH4032*	***PHKG2***	*IOH26832*	***SRPK1***
IOH25914	***ALS2CR7***	*IOH21591*	***DYRK4***	*IOH12327*	***MAPK1***	*IOH21022*	***PIK3C3***	*IOH26788*	***SRPK2***
IOH6107	***ARAF***	*IOH21995*	***EEF2K***	*IOH29657*	***MAPK10***	*IOH11272*	***PIM2***	*IOH38187*	***SRPK2***
IOH21137	***ARAF***	*IOH5879*	***EIF2AK1***	*IOH46259*	***MAPK10***	*IOH39671*	***PIM3***	*IOH21169*	***STK11***
IOH27785	***AURKB***	*IOH29605*	***EIF2AK2***	*IOH11420*	***MAPK11***	*IOH11596*	***PINK1***	*IOH4116*	***STK16***
IOH60102	***AURKC***	*IOH45194*	***EPHA3***	*IOH21127*	***MAPK12***	*IOH21301*	***PKMYT1***	*IOH28976*	***STK16***
IOH22600	***AXL***	*IOH39415*	***EPHB1***	*IOH3435*	***MAPK13***	*IOH26045*	***PKN1***	*IOH56487*	***STK17A***
IOH57009	***BLK***	*IOH28858*	***EPHB3***	*IOH4647*	***MAPK14***	*IOH40845*	***PKN3***	*IOH27608*	***STK24***
IOH27203	***BMP2K***	*IOH29406*	***EPHB4***	*IOH45415*	***MAPK15***	*IOH5070*	***PLK1***	*IOH62982*	***STK24***
IOH22419	***BMPR1A***	*IOH60258*	***ETNK1***	*IOH12559*	***MAPK3***	*IOH57150*	***PLK2***	*IOH27916*	***STK25***
IOH26729	***BMPR1B***	*IOH12251*	***ETNK2***	*IOH29678*	***MAPK6***	*IOH14035*	***PLK3***	*IOH6735*	***STK25***
IOH44883	***BMPR2***	*IOH12546*	***FASTK***	*IOH25729*	***MAPK6***	*IOH21855*	***PNCK***	*IOH9738*	***STK3***
IOH11645	***BMX***	*IOH25748*	***FGFR1***	*IOH5845*	***MAPK7***	*IOH12321*	***PRKAA1***	*IOH29261*	***STK31***
IOH45178	***BRD2***	*IOH21425*	***FGFR1***	*IOH13115*	***MAPK7***	*IOH27141*	***PRKAA1***	*IOH12468*	***STK32A***

IOH14775	BRSK2	IOH36731	FGFR2	IOH29959	MAPK8	IOH26755	PRKAA1	IOH26023	STK32B
IOH39717	BTK	IOH36732	FGFR2	IOH53990	MAPK8	IOH29876	PRKAA2	IOH27515	STK33
IOH21042	BUB1	IOH27231	FGFR2	IOH21974	MAPK9	IOH26286	PRKACA	IOH11118	STK36
IOH45542	C9orf96	IOH13371	FGFR4	IOH42152	MAPK9	IOH10103	PRKACB	IOH13030	STK38
IOH26033	CAMK1D	IOH21152	FGR	IOH3889	MAPKAPK3	IOH27691	PRKACB	IOH11742	STK38L
IOH26807	CAMK2A	IOH42245	FLJ23356	IOH28010	MAPKAPK5	IOH45130	PRKACG	IOH7196	STK4
IOH45432	CAMK2B	IOH60533	FLJ25006	IOH63098	MARK2	IOH29644	PRKCA	IOH6969	STK40
IOH46514	CAMK2D	IOH62170	FLJ25006	IOH10665	MARK3	IOH43969	PRKCB1	IOH40615	STK40
IOH22233	CAMK2G	IOH56137	FRK	IOH29899	MAST1	IOH22061	PRKCB1	IOH29558	STYK1
IOH22804	CAMKK1	IOH21081	FYN	IOH46175	MAST2	IOH26352	PRKCD	IOH4914	SYK
IOH26370	CAMKK1	IOH21890	FYN	IOH4506	MATK	IOH26823	PRKCH	IOH9759	SYK
IOH45150	CAMKK2	IOH3380	GAPDH	IOH45826	MATK	IOH56035	PRKCI	IOH21006	TBK1
IOH37934	CAMKK2	IOH22295	GAPDH (sperm)	IOH36570	MET	IOH42544	PRKCQ	IOH26839	TBK1
IOH12294	CAMKK2	IOH45159	GRK5	IOH46178	MGC16169	IOH5838	PRKCZ	IOH23271	TESK2
IOH21132	CAMKV	IOH14078	GRK6	IOH27623	MGC42105	IOH38433	PRKD2	IOH14312	TNK2
IOH27787	CAMKV	IOH62403	GSG2	IOH5352	MKNK1	IOH40726	PRKG1	IOH62118	TNNI3K
IOH14583	CDC2	IOH11473	GSK3A	IOH37765	MKNK2	IOH45425	PRKX	IOH21149	TRIB2
IOH45143	CDC2L1	IOH4507	GSK3B	IOH53775	MKNK2	IOH59135	PRKY	IOH14340	TRIB3
IOH3095	CDC2L5	IOH14630	HCK	IOH21529	MLKL	IOH45196	PRPF4B	IOH27427	TRIB3
IOH39710	CDK10	IOH27205	HIPK1	IOH35579	MOS	IOH27426	PTK2	IOH27746	TSSK2
IOH41508	CDK10	IOH22334	HIPK1	IOH44055	MYLK	IOH28634	PTK2B	IOH27376	TSSK3
IOH29645	CDK3	IOH22267	HIPK4	IOH45160	MYLK2	IOH37738	PTK6	IOH27738	TTBK2

IOH12184	***CDK4***	*IOH5257*	***HSPB8***	*IOH26905*	***MYO3A***	*IOH14307*	***PXK***	*IOH4197*	***TTK***
IOH5192	***CDK4***	*IOH38087*	***ICK***	*IOH26432*	***NEK10***	*IOH28936*	***RAGE***	*IOH21117*	***TYK2***
IOH4873	***CDK5***	*IOH36576*	***IGF1R***	*IOH28791*	***NEK2***	*IOH5435*	***RET***	*IOH26747*	***TYRO3***
IOH28111	***CDK6***	*IOH6284*	***IKBKB***	*IOH10648*	***NEK3***	*IOH9973*	***RFP***	*IOH28850*	***TYRO3***
IOH3121	***CDK7***	*IOH29703*	***ILK***	*IOH45183*	***NEK4***	*IOH61477*	***RFP***	*IOH14193*	***UHMK1***
IOH2136	***CDK9***	*IOH12099*	***IRAK1***	*IOH14564*	***NEK6***	*IOH40612*	***RIOK1***	*IOH45122*	***ULK3***
IOH25876	***CDKL3***	*IOH29139*	***IRAK3***	*IOH45126*	***NEK7***	*IOH6368*	***RIPK2***	*IOH10843*	***ULK4***
IOH27246	***CDKL5***	*IOH54085*	***ITK***	*IOH45357*	***NEK8***	*IOH26143*	***RIPK3***	*IOH41408*	***VRK1***
IOH5501	***CERK***	*IOH29900*	***JAK2***	*IOH62004*	***NEK9***	*IOH62447*	***RIPK5***	*IOH11957*	***VRK2***
IOH21007	***CLK1***	*IOH61636*	***KIT***	*IOH53723*	***NLK***	*IOH62446*	***RIPK5***	*IOH27845*	***VRK3***
IOH28987	***CLK2***	*IOH34863*	***KSR2***	*IOH63245*	***NPR2***	*IOH39650*	***RP6-213H19.1***	*IOH27052*	***WEE1***
IOH14562	***CLK2***	*IOH45203*	***LATS1***	*IOH4821*	***NRBP1***	*IOH12130*	***RPS6KA1***	*IOH23113*	***WNK1***
IOH6643	***CLK3***	*IOH11874*	***LCK***	*IOH29759*	***NRBP2***	*IOH3648*	***RPS6KA2***	*IOH26758*	***YES1***
IOH4008	***CLK3***	*IOH28609*	***LYK5***	*IOH54159*	***NTRK3***	*IOH63248*	***RPS6KA3***	*IOH22867*	***YSK4***
IOH54018	***CLK4***	*IOH45129*	***LYK5***	*IOH36839*	***NUAK1***	*IOH29892*	***RPS6KA4***	*IOH27465*	***ZAP70***
IOH29870	***CSK***	*IOH39743*	***LYN***	*IOH21129*	***NUAK2***	*IOH45179*	***RPS6KA5***	*IOH28978*	***ZAP70***
IOH59150	***CSNK1A1***	*IOH26189*	***MAK***	*IOH21155*	***OXSR1***	*IOH45431*	***RPS6KA5***	*IOH45475*	***ZAP70***

Table 3.1 The 350 unique kinases on the Kinome microarray. Each distinct IOH number represents a unique protein. More information for each kinase can be found at orf.lifetechnologies.com/cgi-bin/ORF_Browser

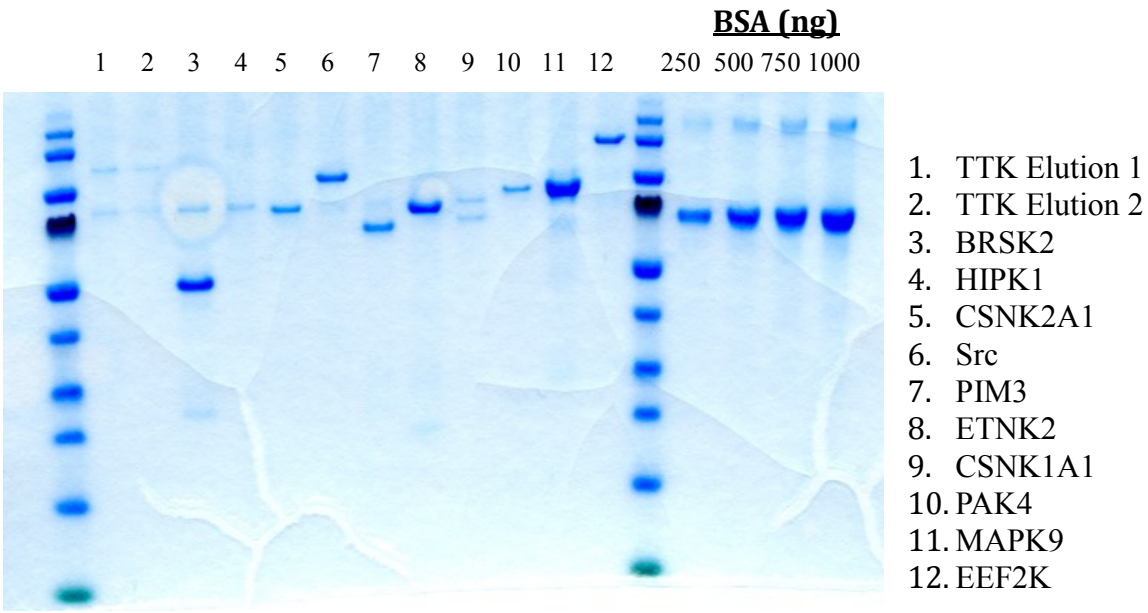

Figure 3.1 Representative coomassie stain for purified GST-tagged kinases (SimplyBlue stain, Life Technologies)

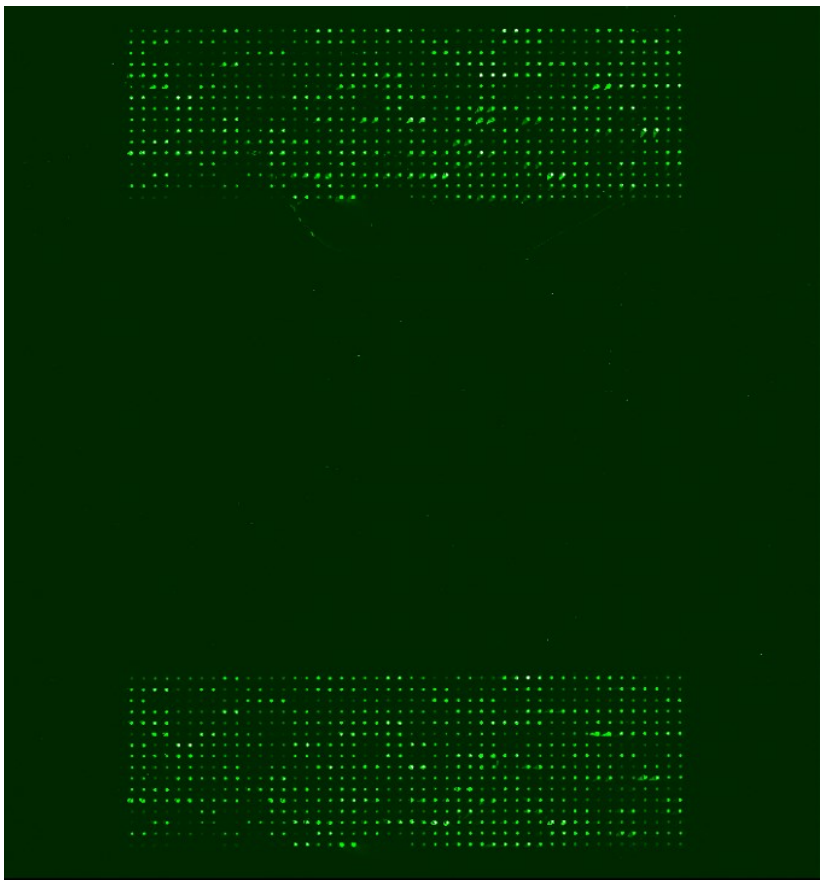

Figure 3.2 The kinome array. High microarray printing quality is shown using anti-GST primary antibody (mouse, Sigma) and goat anti-mouse Cy3 secondary antibody (Life Technologies)

Following purification, the kinases were re-arrayed into 384-well plates before contact printing with a BIO-RAD VersArray ChipWriter Pro System. Each kinase was printed in duplicate as part of a large 16x48 grid that included the 350 kinases as well as a series of controls (including histones, BSA standards, and serial dilutions of GST). Each microarray contained two of these large grids for a total of four replicates of each kinase. Printing quality was assessed using an anti-GST primary antibody and a Cy3-labeled secondary antibody (Figure 3.2). Microarrays used in the antibody-based detection of *O*-linked glycosylation were printed on Full Moon slides (FullMoon Biosystems), while those used in the click chemistry detection were printed on epoxide slides (SuperEpoxy 2, Arrayit Corporation).

3.3 In vitro *Glycosylation Reaction*

In order to assess the breadth of kinome glycosylation, we developed protocols for three different methods of detecting the *O*-GlcNAcylation of the kinome. The first two approaches use antibodies that specifically bind to *O*-glycosylated proteins (RL2 and CTD110.6), while the third approach uses click chemistry to specifically label *O*-glycosylated proteins (Figure 3.3).

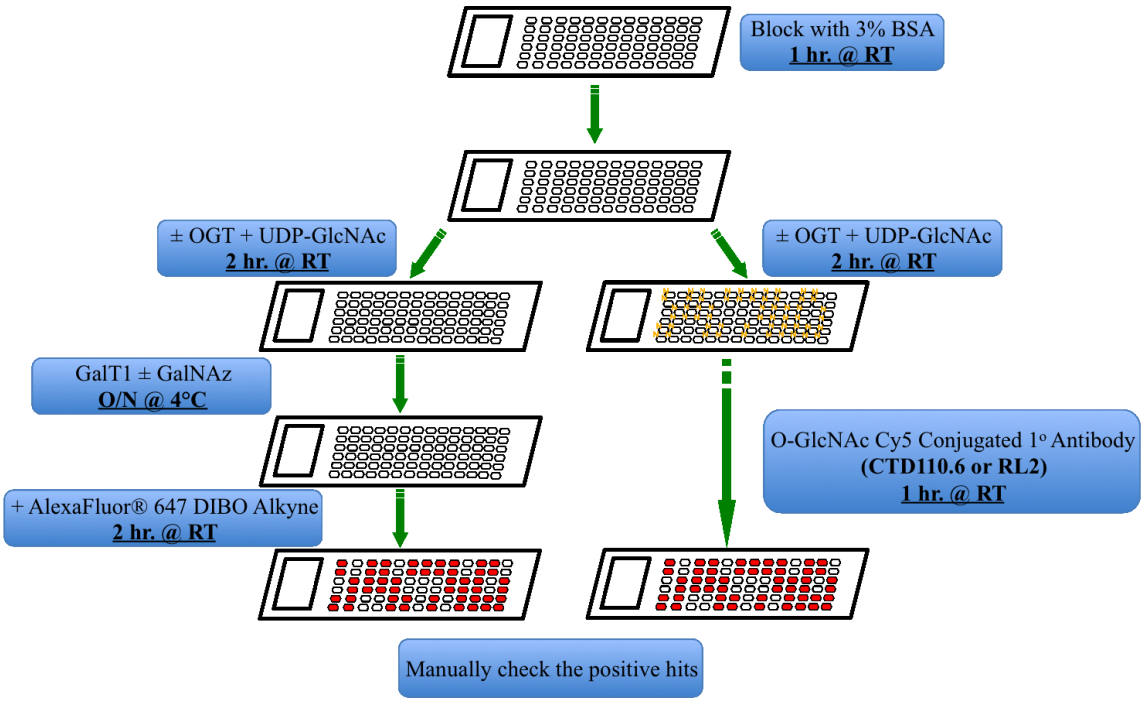

Figure 3.3 General schematic for microarray experiments. Arrays were both labeled with OGT followed by detection with either click chemistry or antibody-based approaches.

The general antibody-based assay involves blocking with 3% BSA in TBST for 1 hour at room temperature with gentle shaking followed by 3 10 minute washes in TBST. Arrays were then treated with ncOGT or enzyme buffer (as negative control) in the presence of UDP-GlcNAc for 2 hours at room temperature under a coverslip and in a humidity chamber. Following OGT labeling, arrays were washed 3 times in TBST (after initial immersion in 400 mL TBST to remove coverslips) for ten minutes before incubation with Cy5-labeled antibody. A final concentration of 20ng/µL was used for RL2 and 5 ng/µL for CTD110.6. Arrays were exposed to each antibody for 2 hours at room temperature under a coverslip and in a humidity chamber. Arrays were quickly immersed in 400 mL TBST and washed 3 times in TBST for ten minutes. Following the final wash, arrays were quickly immersed in preheated 37° ddH_2O and placed in a drying

rack, where they were spun dry by centrifugation (2 minutes at 2000 rpm). Arrays were then scanned with a GenePix 4000 scanner (MDS Analytical Technologies) before careful alignment using GenePix software.

The click chemistry approach used arrays that were printed on epoxide slides (SuperEpoxy 2, Arrayit Corporation). Pilot studies for this approach using FullMoon slides showed tremendous amount of background that was abolished when using the epoxide slides. Arrays were also blocked with SuperBlock (Thermo Scientific) for 1 hour at room temperature. After blocking, arrays were washed and labeled with OGT as before. After labeling (and subsequent wash steps), arrays were further treated with the permissive mutant β-1,4-galactosyltransferase (GalT1 (Y289L)), which transfers azido-modified galactose (GalNAz) from UDP-GalNAz to *O*-GlcNAc residues of modified proteins (Click-iT^{TM} *O*-GlcNAc Enzymatic Labeling System, Invitrogen). GalT1 treatment was performed overnight at 4°C under a coverslip. The following morning, arrays were immersed in 400 mL of TBST to remove the coverslips and washed 3 times for 10 minutes in TBST. Contrary to the antibody approach, the arrays were washed 2 times for 10 minutes with 0.5% SDS as well. This was done to eliminate some background signal that was seen in pilot studies when just performing TBST washes. Following the final wash, the arrays were treated with 1 μM Alexa Fluor® 647 DIBO alkyne for 2 hours at room temperature under a coverslip and in a humidity chamber. This copper-free labeling method had much less background than the traditional copper-catalyzed method. After incubation, arrays were washed 3 times in TBST for 10 minutes before quick immersion in preheated 37° ddH_2O and drying through centrifugation (2000

rpm for 2 minutes). As before, arrays were scanned with a GenePix 4000 scanner (MDS Analytical Technologies) before careful alignment using GenePix software.

Representative images are shown in Figure 3.4 for all three detection methods.

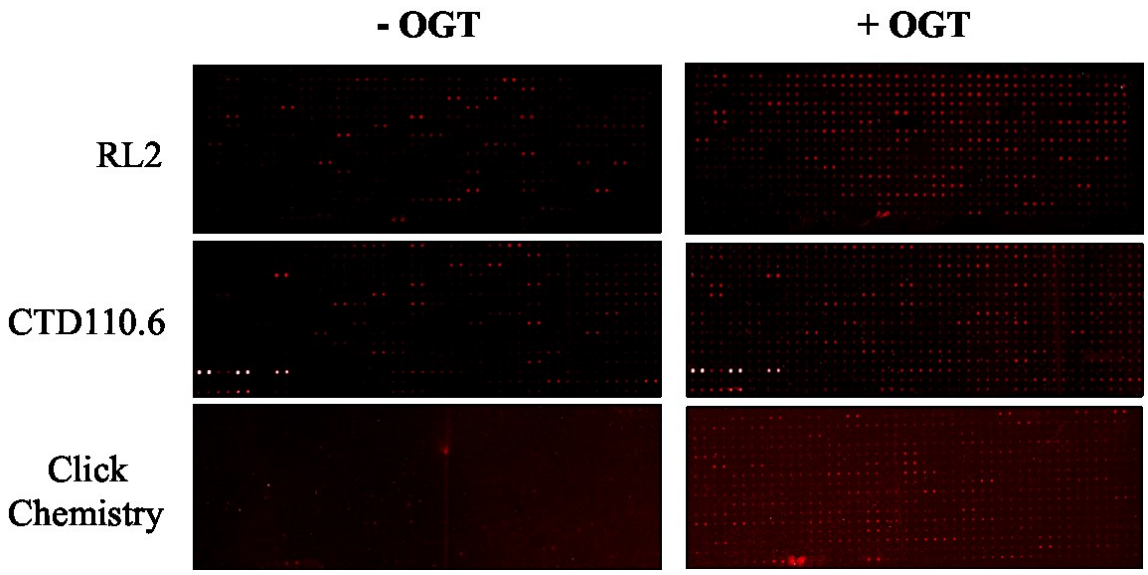

Figure 3.4 Representative images of three detection methods.

In order to identify positive hits, we performed a background subtraction from each spot, averaging the values between replicate spots in the same large block first, and then averaging the values between the two large blocks to obtain an average F-B value for each kinase. The ratio was then determined between OGT treated and untreated arrays to yield a fold increase upon OGT treatment (Figure 3.5).

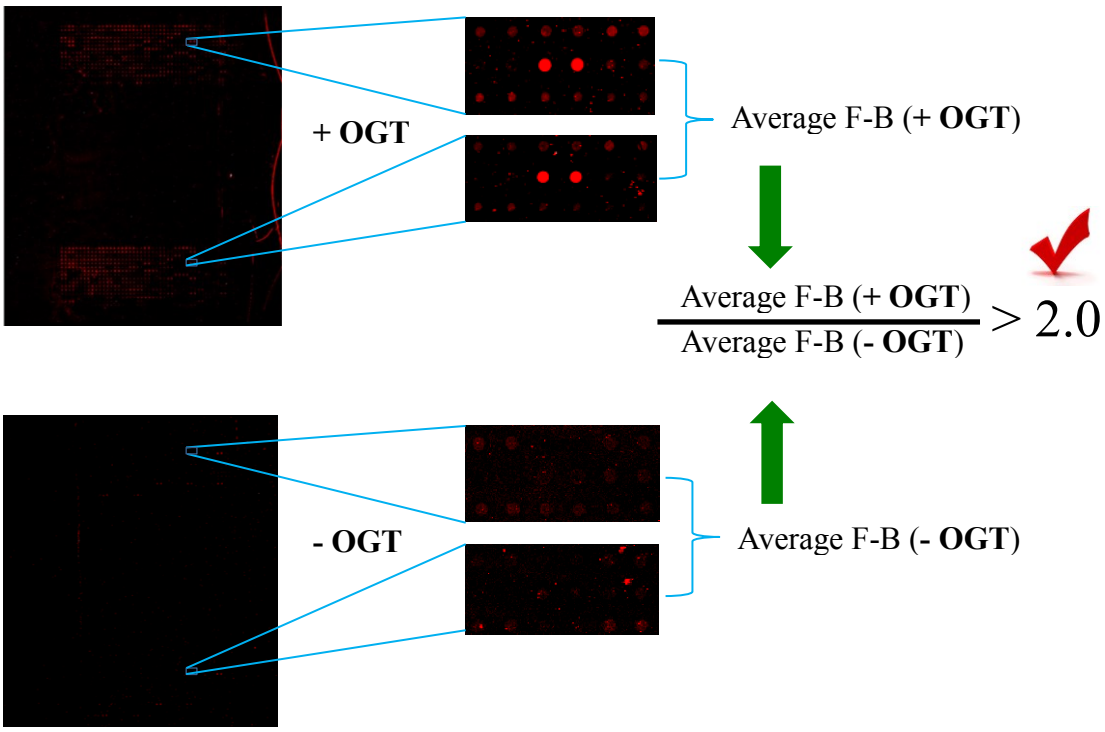

Figure 3.5 Data analysis schematic using RL2 antibody detection as an example. Background subtraction is averaged between replicates in large block as well as between blocks before ratio is determined between OGT treated and untreated samples. Each experiment was repeated twice for each detection method. Threshold was determined experimentally to be 2.0.

As a proof of principle and validation that the antibodies were not non-specifically binding to the array, we diluted each antibody in increasing amounts of free GlcNAc as competition to abolish any non-specific binding. As expected, signal for both RL2 and CTD110.6 decreased dramatically upon competition (Figure 3.6 A). Clostridium perfringens NagJ (CpNagJ) is a close homolog of human *O*-GlcNAcase, the enzyme that hydrolyzes *O*-linked glycosylation. CpNagJ has been shown to have high conservation in the active site with human OGA, and has been shown to have activity on human substrates (Rao et al., 2006). As another proof of principle, we treated the array with or without CpNagJ (and with or without OGT) and saw a dramatic loss in signal. We were then able to rescue some of the hits after CpNagJ treatment using OGT again on the array

(Figure 3.6 B). While we weren't able to rescue all of the signal we had originally, we attribute that to the three consecutive reactions that were performed on the arrays (over the course of 9 hours and 12 washes) which likely led to protein unfolding and an inability of OGT to efficiently modify the proteins after CpNagJ treatment. However, the fact that we saw the expected results for both the competition assay and the CpNagJ treatment gives us great confidence in the hits that we initially generated using both RL2 and CTD110.6.

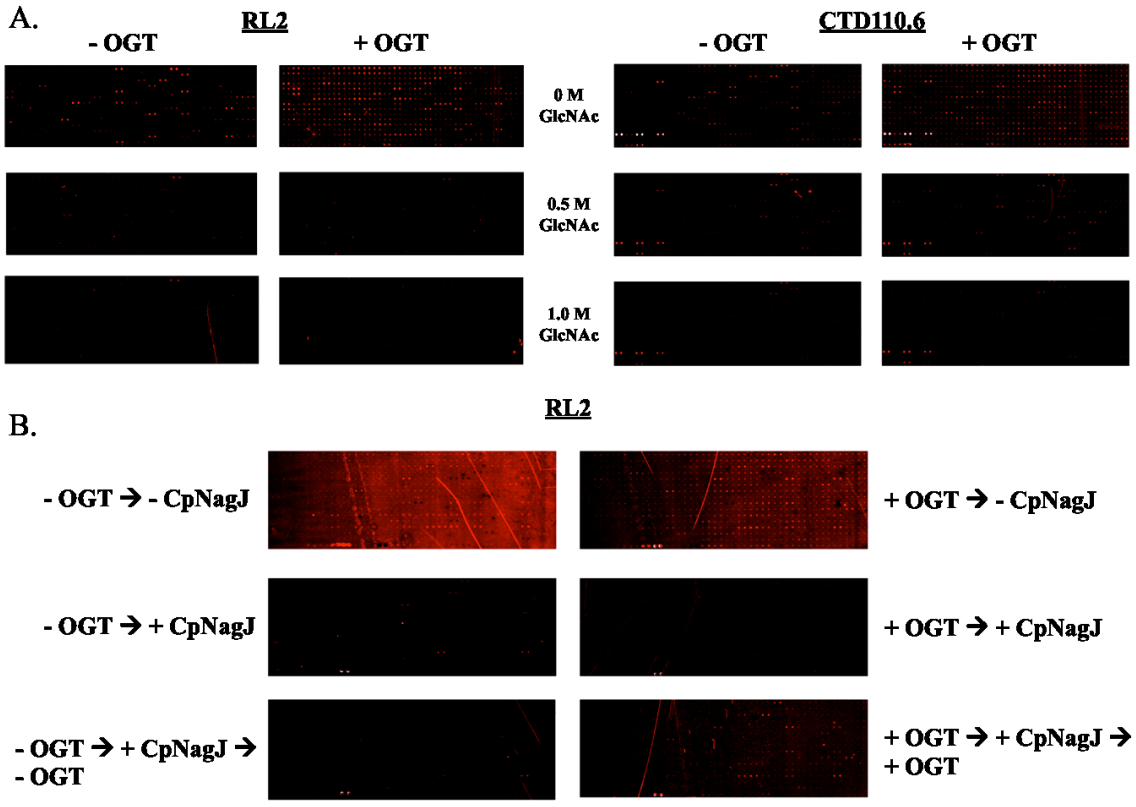

Figure 3.6 Establishing the specificity of RL2 and CTD110.6. (A) By using increasing concentrations of free GlcNAc, we see an abolishment of signal, signifying that the initial signal is a true measure of *O*-GlcNAcylation. (B) Treatment of the arrays with Clostridium perfringens NagJ (CpNagJ) results in a decrease in signal, signifying removal of *O*-GlcNAcylation. The signal is partially rescued upon further treatment with OGT.

After establishing the general validity of our methods using competition and CpNagJ treatment, each detection method was repeated twice, with high reproducibility.

For RL2 detection, 328 of the 341 hits (96%) were shared between duplicate experiments. Similar results were seen for CTD110.6 (218/318; 69%) and click chemistry (183/339; 54%).

3.4 In vitro *Validations*

To begin establishing a threshold for our hit lists, we sought to validate a variety of hits *in vitro*. To that end, several kinases were selected that were in three different categories based on our microarray data; 1) Those that were hit by all three methods in duplicate at a starting threshold of 2.0 (i.e. 2-fold signal intensity above non-treated sample), 2) those that were hit 5 out of 6 times (i.e., were present in all duplicates for each method except for one), and 3) those that were not hit at this initial threshold (Table 3.2).

Hit by All 3 Methods	Hit by 5 of 6 Methods	Not Hit at 2.0
CSNK1A1	BRSK2	PIM3
ETNK2	SYK	CAMKK1 v.1
Src	SRPK2	RIPK3
MAPK8	CSNK2A1	PBK
PAK4	EEF2K	
	TTK	
	MYO3A	
	STK33	
	HIPK1*	

Table 3.2 Hits selected for further validation studies. HIPK1 is a known positive control.

In order to perform *in vitro O*-GlcNAcylation, each kinase was purified as a GST-fusion protein in yeast as before and treated with and without OGT before western transfer. Membranes were either probed with RL2 or CTD110.6 antibody (1:2000 and 1:10,000, respectively) as well as α-FLAG (1:1000) for loading control.

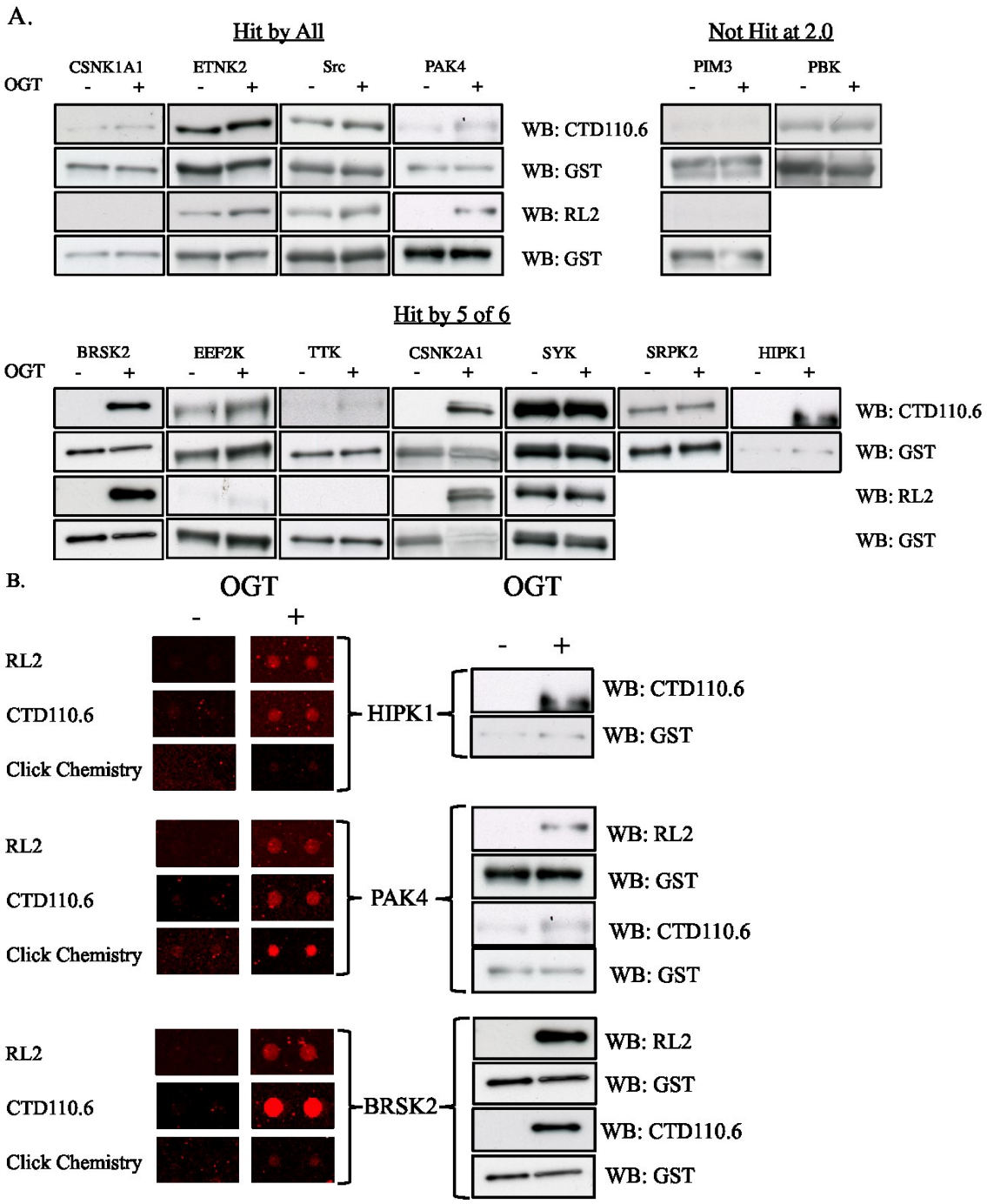

Figure 3.7 First phase validation. (A) *in vitro* validations of GST-fusion kinases were classified three ways; 1) Hit by all methods, 2) Hit by 5 out of 6 replicates, and 3) not hit at 2.0 threshold. (B) Western validations with representative array image references.

As shown in Figure 3.7 A, all four hits that were observed by all three methods were validated (*CSNK1A1, ETNK2, Src,* and *PAK4*) and 5 of 7 hits detected by 5 of the 6

replicates were validated as well (*BRSK2, EEF2K, TTK, CSNK2A1,* and *HIPK1*). Importantly, the two hits tested that did not show signal over thresholds were shown to be negative, as expected (*PIM3* and *PBK*). The high success rate of the *in vitro* validations strengthens our confidence in our findings.

As another method of validation, we performed *in vitro* assays with CpNagJ on 6 proteins (2 from each classification) (Figure 3.8). As expected, treatment with CpNagJ diminished signal and OGT treatment intensified signal among positive hits (*ETNK2, MAPK8, MYO3A,* and *STK33*). No signal was seen for both kinases that were not hits at our threshold of 2.0 (*CAMKK1 v.1* and *RIPK3*).

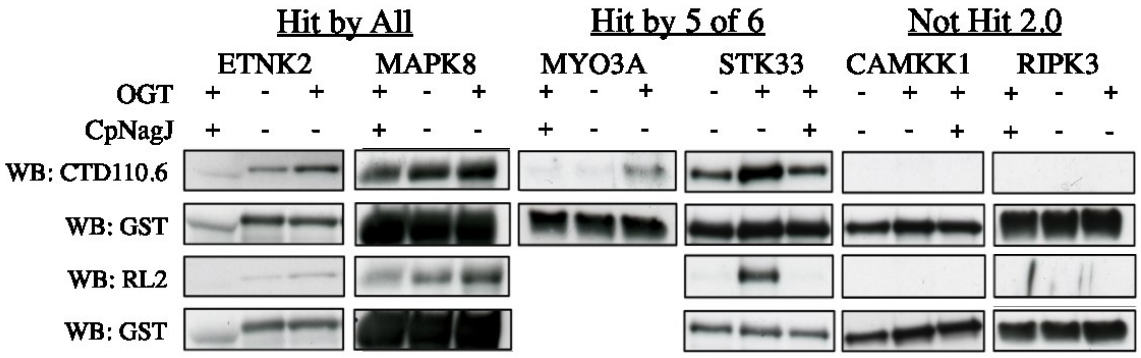

Figure 3.8 *in vitro* validations using CpNagJ to remove *O*-GlcNAcylation. GST-fusion proteins were treated with OGT and/or CpNagJ before western transfer and antibody probing. As expected, CpNagJ treatment diminished signal, while OGT intensified signal among positive hits.

3.5 In vivo *Validations of PAK4 and Other Substrates*

In order to further validate hits, kinases were shuttled into an N-terminal FLAG-tagged mammalian expression vector (pSG5-FLAG, Agilent Technologies) and expressed in HeLa cells. 48 hours after transfection, cells were treated for 4 hours with the *O*-GlcNAcase inhibitor TMG to raise global levels of *O*-GlcNAcylation. Cells were then lysed and treated with anti-FLAG M2 agarose beads (Sigma) overnight at 4°C with

shaking. After washing, proteins were eluted with loading buffer, run on a gel, and transferred to nitrocellulose for western blotting. Blots were probed with both CTD110.6 and anti-FLAG and 9 kinases were validated *in vivo* (Figure 3.9).

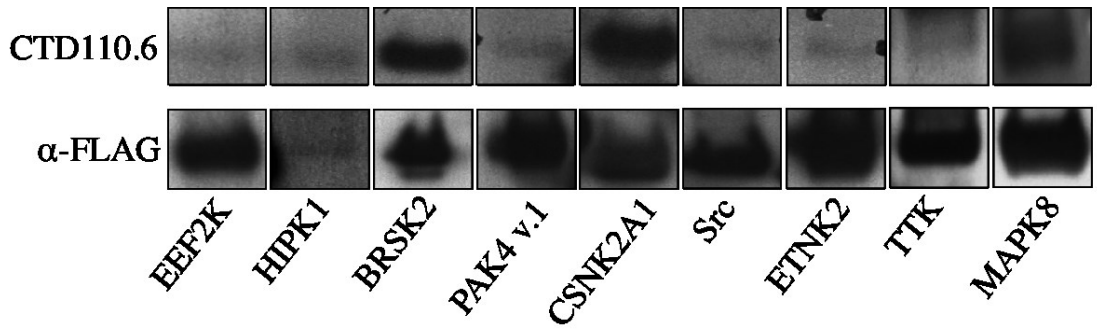

Figure 3.9 *in vivo* validations in HeLa. FLAG-tagged substrates were expressed and 48 hours after transfection, cells were treated with TMG for 4 hours. Cells were lysed and affinity IP was performed using anti-FLAG M2 agarose beads before western transfer.

3.6 *Generation of High Confidence Hit List*

After successfully validating several kinases *in vivo*, we were confident in our initial assessment of a threshold of 2.0, especially since every kinase tested that was shared by all three methods was validated. Therefore, a final high confidence hit list was generated based on two stringent criteria; 1) each hit was present on both replicates using the same detection method and 2) each hit was present using all three detection methods (RL2, CTD110.6, and click chemistry) (Figure 3.10). This high confidence list contains 104 kinases (Table 3.3).

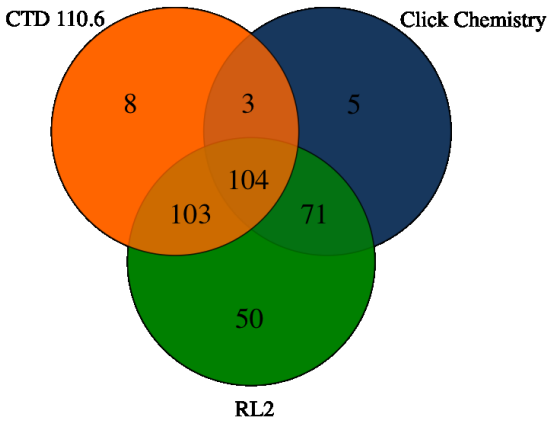

Figure 3.10 Comparison of hits detected by all 3 methods. 104 hits were shared.

Hits By All 3 Methods (RL2, CTD110.6, and Click Chemistry)			
ACVR1	DMPK	MAPK1	PRKCD
ACVR2B	EIF2AK1	MAPK12	PRKD2
ADRBK1	EIF2AK2	MAPK15	PTK6
ADRBK2	EPHA2	MAPK6	RAGE
AKT2	ETNK2	MAPK6	RFP
AKT3 v.1	FASTK	MAPKAPK5	RIPK5
AKT3 v.2	FGR	MARK3	RPS6KB1
ARAF	FLJ25006	MATK	RPS6KB1
AURKB	GAPDH	MET	RPS6KL1
AXL	GSK3A	MOS	SCYL1
BLK	ICK	NEK6	SGK2
BMP2K	IGF1R	NEK7	SRC
BMPR1A	IRAK1	NLK	STK16
BMPR1B	ITK	NRBP2	STK16
BMPR2	KSR2	PAK4 v.1	STK32B
C9orf96	LATS1	PAK4 v.2	STK38
CDK4	LRRK2_2	PCTK3	STK38L
CDK5	LRRK2_3	PCTK3	STYK1
CDK9	LYK5	PDIK1L	SYK
CDKL5	LYN	PHKG2	TRIB3
CSK	MAP2K4	PINK1	TYK2
CSNK1A1	MAP2K4	PKMYT1	ULK3
CSNK1E	MAP2K6	PLK4	VRK1
DCLK1	MAP2K7	PRKAA2	VRK3
DDR1	MAP3K7	PRKACG	YES1
DDR2	MAP4K5	PRKCB1	ZAP70

Table 3.3 List of 104 kinases shared between 3 detection methods.

In order to begin characterization of the kinases, we performed gene ontology analysis to see if any terms were enriched among our dataset. This was done using the online software Gene Ontology Enrichment Analysis Software Toolkit (GOEAST; Zheng et al., 2008). As expected, there weren't many significant GO enrichments for a dataset consisting entirely of kinases. Almost all of the GO terms were related to generic kinase function (i.e. ATP binding, protein S/T kinase activity, enzyme binding, etc.). However, some interesting terms were seen that have been shown to be involved with OGT as well (Table 3.4). Given the fact that there is only one OGT gene in all advanced organisms, it is not surprising that there are a wide variety of processes that may be involved with OGT, especially if our observation is that a large percentage of the kinome is *O*-GlcNAcylated.

GOID	Ontology	Term	# in Dataset	Log Odds Ratio	p-value
GO:0006950	Biological process	Response to stress	**44**	2.232	1.75E-17
GO:0043408	Biological process	Regulation of MAPK cascade	**22**	3.884	2.49E-17
GO:0002757	Biological process	Immune response-activating signal transduction	**18**	4.140	3.12E-15
GO:0008219	Biological process	Cell death	**23**	2.917	4.57E-12
GO:0006915	Biological process	Apoptotic process	**20**	2.962	1.34E-10
GO:0097472	Molecular function	Cyclin-dependent protein kinase activity	**7**	6.261	6.40E-10
GO:0016477	Biological process	Cell migration	**17**	3.105	1.68E-09
GO:0045595	Biological process	Regulation of cell differentiation	**21**	2.513	9.98E-09
GO:0007596	Biological process	Blood coagulation	**12**	3.396	1.89E-07
GO:0042060	Biological process	Wound healing	**13**	3.047	6.62E-07
GO:0045121	Cellular component	Membrane raft	**9**	3.783	2.08E-06
GO:0032868	Biological process	Response to insulin stimulus	**9**	3.599	5.74E-06
GO:0008283	Biological process	Cell proliferation	**11**	2.545	0.000210
GO:0005815	Cellular component	Microtubule organizing center	**9**	2.599	0.001116

Table 3.4 Gene ontology analysis from GOEAST. Most enrichment categories were related to general kinase function, but several interesting enrichments are seen. The larger the log odds ratio (base 2), the stronger the enrichment. P-values are calculated hypergeometrically to calculate enrichment.

3.7 Identification and Confirmation of Glycosylation Sites Using Mass Spectrometry

Next, we set to identify actual sites of *O*-GlcNAcylation for some of our kinase hits. To that end, BRSK2 and PAK4, which had both been validated *in vitro* and *in vivo*, were purified as GST-fusion proteins in yeast as before. They were then labeled and submitted for mass spectrometry. S691 of BRSK2 was identified as being glycosylated (Figure 3.11 A), while PAK4 showed 4 potential glycosylation sites (S12, S242, S243, and S267) (Figure 3.11 B).

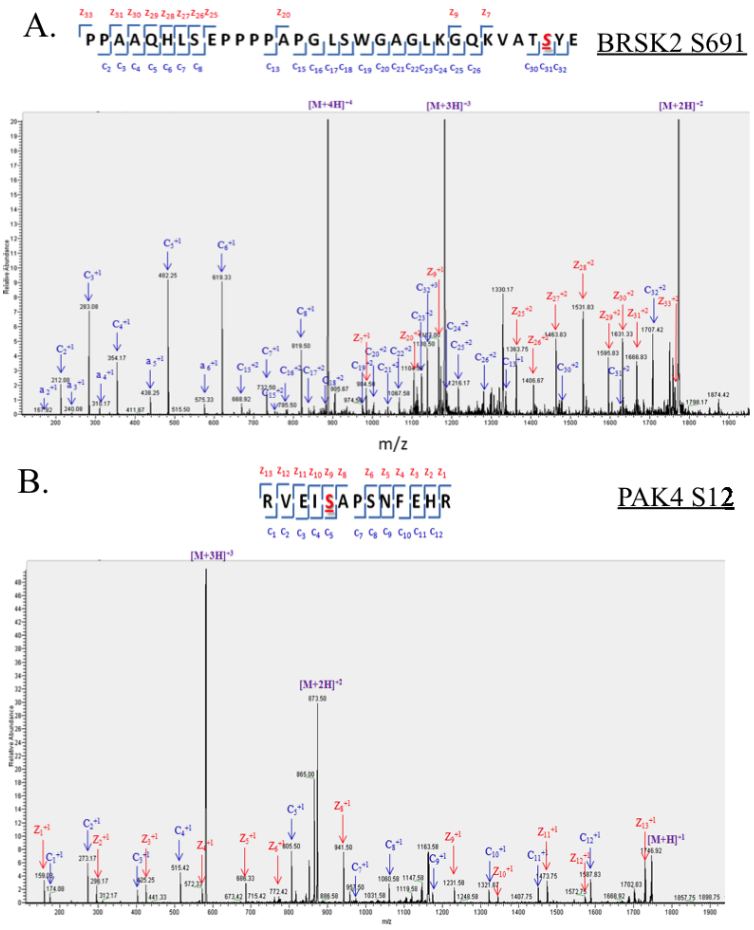

Figure 3.11 Spectra from ETD mass spectrometry (A) S691 of BRSK2 identified and (B) S12 of PAK4 identified as *O*-GlcNAcylated. S242, 243, and 267 of PAK4 were also identified (not shown).

In order to validate these sites, we performed S→A mutagenesis on PAK4 at each site individually, as well as a 4A mutant containing all four putative sites. Following mutagenesis, entry clones were shuttled into the pSG5-FLAG expression vector for expression in mammalian cells. As before, cells were transfected with PAK4 for 48 hours before 4 hours of TMG treatment. Cells were then lysed and immunoprecipitated with anti-FLAG M2 agarose beads overnight before washing and separation by gel electrophoresis. Proteins were then transferred to nitrocellulose for western blotting using both CTD110.6 (1:10,000) and α-FLAG (1:1000) antibodies. As seen in Figure 3.12, the S12A mutant exhibits an approximately 50% decrease in intensity compared to wild type,

and the 4A mutant (containing S12A, S242A, S243A, and S267A) shows >80% reduction. These results signify that the sites identified by mass spectrometry are of high quality. More importantly, it provides us with great confidence that the hits we obtained on our microarray have enough fidelity to be able to successfully identify individual sites of *O*-GlcNAcylation.

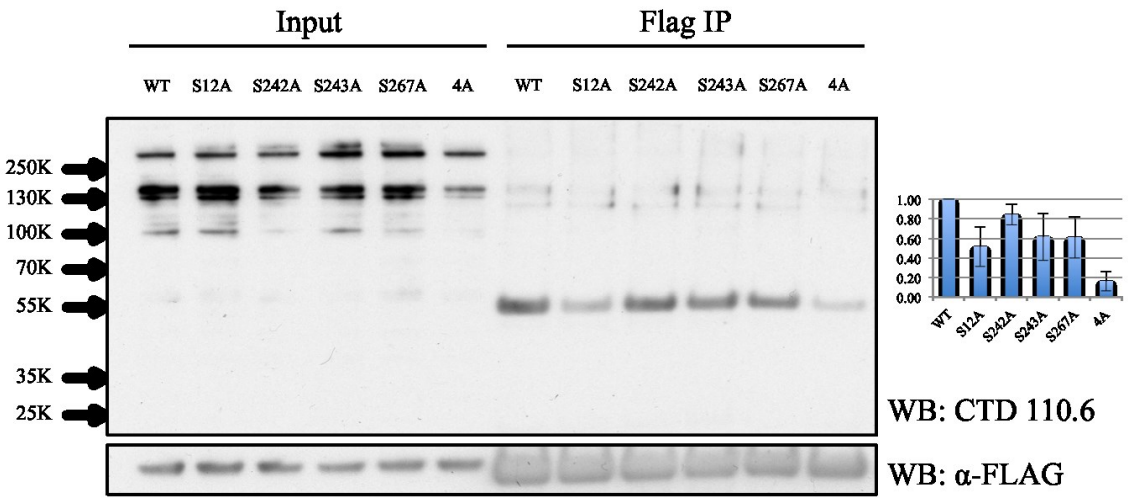

Figure 3.12 *in vivo* mutagenesis studies validate identified *O*-GlcNAcylated residues in PAK4. Insert: N=5

3.8 Discussion

O-linked *N*-acetylglucosamine (*O*-GlcNAc) was discovered in the early 1980's when bovine milk galactosyltransferase (GalT1) was used to probe for terminal *N*-acetylglucosamine moieties of glycoconjugates in living cells, yet it continues to be difficult to detect and quantify in many systems (Torres et al., 1984). Part of the reason for this is the fact that *O*-GlcNAcylation is generally undetected by standard analytical protein methods, such as gel electrophoresis and HPLC (Roquemore et al., 1994; Zachara et al., 2004). For one, *O*-GlcNAcylation generally does not result in gel shifts during electrophoresis, isoelectric focusing, or even 2D gels. Secondly, the sugar moieties can be

readily hydrolyzed by cellular hexoaminidases during cell damage or protein purification. Third, traditional mass spectrometry renders the *O*-GlcNAc modification very labile, and is traditionally lost (Wang et al., 2010). This, combined with the fact that *O*-GlcNAc-peptide ion signals are generally suppressed even when modified peptides are in the majority, results in a difficult PTM to study in most conventional labs. Therefore, it is of great interest to develop more sensitive and global assays to be able to more fully elucidate the breadth of the *O*-GlcNAcylation of the proteome.

Kinases represent an intriguing group of proteins to first study as their cross-talk with OGT and OGA have been shown to be extensive (Wang et al., 2008). Inhibition of GSK3β has been shown to increase *O*-GlcNAcylation of many cytoskeletal and heat shock proteins, and decrease in transcription factors and RNA-binding proteins (Wang et al., 2007). Dysregulation of *O*-GlcNAcylation/phosphorylation appears to be important in the pathologies of diabetes and Alzheimer's, as well as certain cancers (Dias et al., 2007, Kawauchi et al., 2009). Because of this, have a greater understanding of how the kinome is affected by OGT is of great importance.

The difficulty in detection of *O*-GlcNAcylated proteins, coupled with the obvious importance of the modification on the kinome, led us to develop a novel way to quickly and specifically identify OGT substrates. Here, we show that through the use of both antibody and click chemistry techniques on kinome arrays, we were able to generate a high confidence list of *O*-GlcNAcylated kinases.

Utilizing the two GlcNAc detection antibodies RL2 and CTD110.6 allowed us to detect *in vitro O*-GlcNAcylation on dozens of kinases. The specificity of our assay was

exhibited through competition using free GlcNAc during antibody binding, causing a dramatic decrease in observed signal. Also, treatment with the OGA homolog Clostridium perfringens NagJ (CpNagJ) abolished signal as well, further validating our method. Treatment of arrays with OGT after CpNagJ treatment was also able to rescue many of the hits we initially observed.

A novel click chemistry approach was applied to the microarrays, which concurrently labeled *O*-GlcNAc residues with an additional azide sugar moiety that could react specifically with a copper-free fluorescent dye (Figure 3.4). Combining hits from all three methods allowed us to arrive at a high confidence hit list containing 104 kinases that were all hit in 2 replicates using each detection method.

Several kinases were validated *in vitro* which allowed us to establish a threshold for our analysis. Among the hits shared by all three methods, 5 of 5 were validated *in vitro*. Similarly, among those with a less stringent cutoff (hit by 5 of the 6 assays performed), 7 of 9 were validated *in vitro*. Importantly, among the 4 hits below our threshold, none of them showed any signal in our *in vitro* assays (Figures 3.7 and 3.8). These results give us great confidence in the fidelity of our assay and points to a large contingent of the kinome (104/350; 30% of those tested using strict criteria) as being *O*-GlcNAcylated.

Gene Ontology analysis showed expected results given all hits were kinases, but also shared some characteristics with OGT including apoptosis, insulin signaling, and microtubule organization. Many other terms were enriched varying from stress and immune response to cell migration and wound healing. Given the apparent depth and

variety of kinases in our final list, this is not surprising, and further studies are necessary to tease out this observation.

Further validations were carried out in HeLa cells and all proteins tested showed *O*-GlcNAcylation after TMG treatment and FLAG immunoprecipitation (Figure 3.9). We were also able to identify *O*-GlyNAcylation sites on two proteins, *BRSK2* (at S691) and *PAK4* (at S12, S242, S243, and S267). Mutagenesis was performed on *PAK4*, and FLAG pulldown confirmed the identity of these sites. Therefore, we are very confident in the list that we have generated regarding the *O*-GlcNAcylation of the kinome as we have gone from chip, to test tube, to cellular, to site-specific identification of *O*-GlcNAcylation. Importantly, our final list was arrived using very strict criteria, as each hit had to be identified a total of 6 times across 3 detection methods. We were also able to validate a large percentage of kinases that showed signal even when only 5 of 6 arrays showed signal. Therefore, although our threshold is harsh, many other kinases that didn't make the final cut indeed may (and are, as is the case with *BRSK2*) still be *O*-GlcNAcylated.

This study represents a promising jumping off point for both the field of *O*-GlcNAcylation and proteomics, as we have established a novel 3-pronged, high sensitivity microarray approach for the identification of *O*-GlcNAcylated proteins on protein microarrays. Further studies will be able to expand the scope of the assay to more than just the kinome, further expanding the breadth of OGT in biology.

3.9 Materials and Methods

Protein Microarray Fabrication

Kinases were cultured in a high throughput manner (16 mL cultures) and purified as described before (Jeong et al., 2012) but also in the presence of Phosphatase Inhibitor Cocktails 2 and 3 (Sigma) to maintain kinase activity. After elution, proteins were re-arrayed into a 384-well plate and printed using a split pin on a BIO-RAD VersArray ChipWriter Pro System.

In vitro O-GlcNAcylation Assay

10x Calf Intestinal Alkaline Phosphatase (CIP; New England Biolabs, M0290S) was diluted into either 50 mM HEPES (pH 7.5) or 50 mM Tris-HCl (pH7.5) to make a 1x solution. Reaction mix was made by mixing 1μL (1 Unit) of 1x CIP, 1 μg ncOGT (up to 3 μg), 1 μL of 250 nM 5'-AMP (SIGMA, A2252), 1 μL of 200 mM UDP-GlcNAc (10 mM Final Concentration; SIGMA, U4375) diluted up to 120 μL/array. The negative controls contained everything but the ncOGT. Arrays were blocked for 1 hour at room temperature in 3% BSA in TBST and washed 3 times for 10 minutes in TBST. Excess buffer was wicked off of the arrays onto a kimwipe. 120 μL of reaction mix was added to each array and a coverslip was carefully placed on top of the slide, making sure to avoid bubbles. Arrays were placed in a humidity chamber for 2 hours at room temperature. After labeling, the arrays were quickly immersed in 400 mL of TBST to remove the coverslips and were washed 3 times for 10 minutes in TBST before the next step.

This same general protocol was used for the validation studies in test tubes.

Antibody Binding Protocol (RL2 and CTD110.6)

After OGT labeling and washes, microarrays were treated with 120 μL of antibody solution diluted in 3% BSA in TBST. Both Cy5-RL2 and Cy5-CTD110.6 were purified in the Hart lab for use on all of these experiments. For RL2 binding, a final concentration of 20 ng/μL was used. For CTD110.6 binding, a final concentration of 5 ng/μL was used.

Click Chemistry Labeling Protocol

Note: slides used for click chemistry experiments were epoxide slides, not FullMoon slides. Blocking was also performed using SuperBlock (Thermo) for 1 hour at room temperature. After OGT labeling and washes, arrays were labeled with the azide sugar GalNaz by GalT1 from the Click-iT™ *O*-GlcNAc Enzymatic Labeling System (Invitrogen, C33368) overnight at 4° C. After wash steps, arrays were incubated with Alexa Fluor® 647 DIBO alkyne (Invitrogen, C10408) for 2 hours at room temperature.

CpNagJ Treatment

CpNAgJ was diluted to a final concentration of 50ng/μL in TBS before the addition of 120 μL to each array followed by coverslip addition and placement in a humidity chamber. Arrays were incubated at 37° C for 2 hours before washes and treatment with antibody. Purified CpNAgJ was generously provided by the van Aalten lab (Rao et al., 2006).

The same general protocol was used for validation reactions in test tubes.

Mass Spectrometry Preparation

Recombinant proteins (GST-PAK4 or GST-BRSK2, 2μg) were purified from yeast as GST-fusion proteins. The eluates were then diluted with 2μL 500mM HEPES (pH 7.5) to 50mM HEPES (pH 7.5). One Unit of Alkaline Phosphatase CIP, 1μg ncOGT, and 1μL of 200 mM UDP-GlcNAc were added to solution to make the final volume of 20 μl. The recombinant protein was *in-vitro* labeled with ncOGT for 2 hours at room temperature. After the reaction, the solution was further diluted with 6μL of 4x laemmli (gel loading) buffer and separated on 10% Tris-HCl Precast Gel (Bio-rad) before staining with Coomassie Bright Blue R-250. Individual gel bands were cut out and destained with 100 mM ammonium bicarbonate buffer, Acetonitrile (ACN) in buffer (1:1), and 100% ACN. The samples were dried, reduced with 10 mM DTT at 37 ^{o}C for 2 hours, and then incubated with 20 mM IAA at room temperature in darkness for 45 min. Protein were trypsin-digested overnight (40 ng) at 37^{o}C in 25 mM ammonium bicarbonate. Peptides were sequentially extracted with 5% FA in 50% ACN twice, and finally with 100% ACN with sonication at each stage. The extracts from each individual gel band were pooled and dried by a vacuum concentrator. The dried peptide sample was dissolved in 0.1 % TFA and the sample was desalted with a C18 spin column (The Nest Group, Inc.).

The desalted sample was analyzed using an LTQ-Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) with electron transfer dissociation (ETD). The peptides in the sample were separated by C18 reverse phase column with nano-flow HPLC, then analyzed by LTQ-Orbitrap Elite mass spectrometer using a spray voltage of 2 kV. The full MS scans were acquired in the FT analyzer with the following parameters:

resolution 60,000; mass scan range (m/z) 400.0-1800.0. The 7 HCD MS^2 were performed on the top 7 ions in full MS, meanwhile, the ETD MS^2 was triggered with the same parent ions as when the product ions (204.0866, O-GlcNAc oxonium ionscans) were detected in the HCD MS^2. For both HCD and ETD MS^2, Isolation Width: 3.00; Activation Time is 130.00 for ETD MS^2 and 0.100 for HCD MS^2. For the Dynamic exclusion, Repeat Count, 1; Repeat Duration, 30.00; Exclusion List Size, 500; Exclusion Duration, 80.00.

The MS2 spectra were searched against a modified target-decoy database including the sequences of the individual target proteins and a decoy version of a small inhomogeneous protein database (Anal Chem. 2009, 81(14): 5794-805) using MASCOT with the following parameters: enzyme, trypsin (KR/P); enzyme limits, fully enzymatic (cleaves at both ends); precursor-ion mass tolerance, 50 ppm; fragment-ion mass tolerance, 0.8 Da; missed cleavages, 2; fixed modification, Carbamidomethyl (C); Variable modifications, HexNAc (S), HexNAc (T), Oxidation (M). All potential O-GlcNAcylated peptide MS^2 spectra were carefully manually validated to localize the O-GlcNAcylation site.

Cell Culturing, Pulldown, and Western Blots

HeLa cells were grown in a 6-well culture dish in standard DMEM with 10% FBS and Pen/Strep to approximately 85% confluency before transfection of 850 ng/kinase in pSG5-FLAG mammalian expression vector using PEI (2μL PEI/μg DNA). Media exchange was performed 24 hours after transfection. Cells were treated with 1μM TMG for 4 hours 48 hours after transfection before lysis in cold lysis buffer (125 μL/well).

Lysis Buffer components: 50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1mM EDTA, 1% Triton X-100, cOmplete ULTRA EDTA-free protease inhibitor (Roche), with 1μM TMG or 50 mM GlcNAc. 375 μL of each lysate was incubated with 20 μL of ANTI-FLAG® M2 Agarose Beads (Sigma) overnight at 4° C on a rotator. The next morning, beads were washed 4x with 300 μL cold lysis buffer before elution with 2x loading buffer. Samples were boiled for 5 minutes before gel electrophoresis and wet western transfer. Membranes were blocked for 1 hour at room temperature with 5% milk in TBST and washed three times with TBST before overnight incubation with primary antibody (RL2 – 1:2000; CTD110.6 – 1:10,000) diluted in 3% BSA in TBST. Blots were then washed with TBST three times before incubation with secondary HRP antibody diluted in 3% BSA in TBST.

3.10 References

Comer, F.I., and Hart, G.W. (2001). Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II. Biochemistry *40,* 7845-7852.

Dias, W.B., Cheung, W.D., and Hart, G.W. (2012). O-GlcNAcylation of kinases. Biochem. Biophys. Res. Commun. *422,* 224-228.

Dias, W.B., Cheung, W.D., Wang, Z., and Hart, G.W. (2009). Regulation of calcium/calmodulin-dependent kinase IV by O-GlcNAc modification. J. Biol. Chem. *284,* 21327-21337.

Dias, W.B., and Hart, G.W. (2007). O-GlcNAc modification in diabetes and Alzheimer's disease. Mol. Biosyst *3,* 766-772.

Du, X.L., Edelstein, D., Dimmeler, S., Ju, Q., Sui, C., and Brownlee, M. (2001). Hyperglycemia inhibits endothelial nitric oxide synthase activity by posttranslational modification at the Akt site. J. Clin. Invest. *108,* 1341-1348.

Hart, G.W., Housley, M.P., and Slawson, C. (2007). Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. Nature *446,* 1017-1022.

Hart, G.W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. (2011). Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. Annu. Rev. Biochem. *80,* 825-858.

Jeong, J.S., Jiang, L., Albino, E., Marrero, J., Rho, H.S., Hu, J., Hu, S., Vera, C., Bayron-Poueymiroy, D., Rivera-Pacheco, Z.A.*, et al.* (2012). Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. Mol. Cell. Proteomics *11,* O111.016253.

Kamemura, K., Hayes, B.K., Comer, F.I., and Hart, G.W. (2002). Dynamic interplay between O-glycosylation and O-phosphorylation of nucleocytoplasmic proteins: alternative glycosylation/phosphorylation of THR-58, a known mutational hot spot of c-Myc in lymphomas, is regulated by mitogens. J. Biol. Chem. *277,* 19229-19235.

Kawauchi, K., Araki, K., Tobiume, K., and Tanaka, N. (2009). Loss of p53 enhances catalytic activity of IKKbeta through O-linked beta-N-acetyl glucosamine modification. Proc. Natl. Acad. Sci. U. S. A. *106,* 3431-3436.

Rao, F.V., Dorfmueller, H.C., Villa, F., Allwood, M., Eggleston, I.M., and van Aalten, D.M. (2006). Structural insights into the mechanism and inhibition of eukaryotic O-GlcNAc hydrolysis. Embo j. *25,* 1569-1578.

Roquemore, E.P., Chou, T.Y., and Hart, G.W. (1994). Detection of O-linked N-acetylglucosamine (O-GlcNAc) on cytoplasmic and nuclear proteins. Methods Enzymol. *230,* 443-460.

Slawson, C., Copeland, R.J., and Hart, G.W. (2010). O-GlcNAc signaling: a metabolic link between diabetes and cancer? Trends Biochem. Sci. *35,* 547-555.

Song, M., Kim, H.S., Park, J.M., Kim, S.H., Kim, I.H., Ryu, S.H., and Suh, P.G. (2008). o-GlcNAc transferase is activated by CaMKIV-dependent phosphorylation under potassium chloride-induced depolarization in NG-108-15 cells. Cell. Signal. *20,* 94-104.

Torres, C.R., and Hart, G.W. (1984). Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. J. Biol. Chem. *259,* 3308-3317.

Wang, Z., Gucek, M., and Hart, G.W. (2008). Cross-talk between GlcNAcylation and phosphorylation: site-specific phosphorylation dynamics in response to globally elevated O-GlcNAc. Proc. Natl. Acad. Sci. U. S. A. *105,* 13793-13798.

Wang, Z., Pandey, A., and Hart, G.W. (2007). Dynamic interplay between O-linked N-acetylglucosaminylation and glycogen synthase kinase-3-dependent phosphorylation. Mol. Cell. Proteomics *6,* 1365-1379.

Wang, Z., Udeshi, N.D., O'Malley, M., Shabanowitz, J., Hunt, D.F., and Hart, G.W. (2010). Enrichment and site mapping of O-linked N-acetylglucosamine by a combination of chemical/enzymatic tagging, photochemical cleavage, and electron transfer dissociation mass spectrometry. Mol. Cell. Proteomics *9,* 153-160.

Yang, W.H., Kim, J.E., Nam, H.W., Ju, J.W., Kim, H.S., Kim, Y.S., and Cho, J.W. (2006). Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. Nat. Cell Biol. *8,* 1074-1083.

Zachara, N.E., Cheung, W.D., and Hart, G.W. (2004). Nucleocytoplasmic glycosylation, O-GlcNAc: identification and site mapping. Methods Mol. Biol. *284,* 175-194.

Zeidan, Q., and Hart, G.W. (2010). The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. J. Cell. Sci. *123,* 13-22.

Zheng, Q., and Wang, X.J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. *36,* W358-63.

Chapter 4: Outlook and Future Directions of Protein Microarray Technology

4.1 Future Perspective

The development of the protein microarray over the past decade has led to large strides in our scientific knowledge. Combining versatility and specificity with the high throughput nature of the technology allows many labs to study anything from protein binding interactions, to post-translational modifications, to biomarkers quickly and reproducibly.

While some studies have identified protein modifications using MS/MS approaches and even protein microarrays (Newman et al., 2013), there is still a great deal left unknown about how an enzyme specifically targets its substrate. Peptide arrays are one growing field with a promising outlook on motif identification. Studies by Rothbart and Mah have utilized peptide arrays to identify histone modifications and PTM motifs (Rothbart et al., 2012; Mah et al., 2006). The fabrication techniques still leave room for improvement, however, as using *in vitro* translated peptides is expensive and slow, and *in situ* array formation lacks in quality control. *In situ* arrays have the most straightforward approach to improvement, but this subset of microarrays will still be a powerful tool for modification identification in the future.

Although useful, to date, many studies have used microarrays with limited proteomic coverage, limiting the characterization of what is truly a "proteome" array. This is especially true in higher eukaryotes. The Zhu laboratory has taken great strides in increasing this coverage with their human proteome array, which consists of over 17,000

full-length human proteins, and is the largest to date (Jeong et al., 2012). Utilizing this array could lead to further understanding and deeper characterization of the many PTM that have been identified via MS/MS approaches. It is believed that these technologies will widely be used in the identification of biomarkers as well. Recent studies have focused more on infectious and autoimmune diseases, but can be greatly useful in the identification of cancer biomarkers, potentially leading to diagnostic and prognostic capabilities. The human proteome array is the ideal array to use for this technique, as it currently is the most comprehensive way to screen for antigens. Although large screens can be prohibitively expensive, using the two-phase strategy in the AIH study described earlier (Song et al., 2010) could be the approach used clinically to obtain the greatest results. Using mammalian cell lysates on these arrays could also provide a more relevant pseudo-*in vivo* system, building upon the technique utilized by Merbl and Kirschner (2009). The benefit of having reaction cofactors in the reaction mix cannot be overlooked, which should drastically reduce the number of false positives and false negatives that can plague typical studies.

Perhaps the most exciting new application in the study of microarrays is the advent of SPR and OI-RD technologies. While still in its infancy, many experiments can be imagined that utilize this technique. One of the largest hurdles in using antibody arrays and in the identification of biomarkers is the dearth of specific commercial antibodies. To date, there is no standard metric for antibody affinities used for commercially available antibodies, which always results in a series of optimization steps by each lab in order to determine the best conditions for a particular assay. Using SPR/OI-RD approaches could easily be used to screen antibody affinities, and in turn,

create some sort of unified measurement of affinity. These label-free approaches could also theoretically be used for large-scale drug screens, without compromising the binding efficiency through conjugation with fluorophores or other large molecules such as biotin. One of the largest benefits of these approaches, however, is the ability to measure these interactions in real-time. That, coupled with the high throughput nature of microarrays, can provide a powerful system of measuring enzyme kinetics and dynamics, drastically increasing our knowledge of these interactions.

While much work is still to be done in truly understanding our world, protein microarrays appear up to the task for large-scale proteomic studies. No other system allows for the amount of sensitivity, versatility, and high throughput ability than a microarray. It is our belief that the future of personalized medicine will also utilize this technology in one form or another, leading to better diagnosis and prognosis in many diseases. As this technology continues to mature, we can look forward to all the new connections that will be made, for protein microarrays are truly a versatile tool for scientific discovery.

4.2 - References

Hu, J., Rho, H.S., Newman, R.H., Hwang, W., Neiswinger, J., Zhu, H., Zhang, J., and Qian, J. (2013). Global analysis of phosphorylation networks in humans. Biochim. Biophys. Acta

Jeong, J.S., Jiang, L., Albino, E., Marrero, J., Rho, H.S., Hu, J., Hu, S., Vera, C., Bayron-Poueymiroy, D., Rivera-Pacheco, Z.A.*, et al.* (2012). Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. Mol. Cell. Proteomics *11,* O111.016253.

Mah, A.S., Elia, A.E., Devgan, G., Ptacek, J., Schutkowski, M., Snyder, M., Yaffe, M.B., and Deshaies, R.J. (2005). Substrate specificity analysis of protein kinase complex Dbf2-Mob1 by peptide library and proteome array screening. BMC Biochem. *6,* 22.

Merbl, Y., and Kirschner, M.W. (2009). Large-scale detection of ubiquitination substrates using cell extracts and protein microarrays. Proc. Natl. Acad. Sci. U. S. A. *106,* 2543-2548.

Rothbart, S.B., Krajewski, K., Strahl, B.D., and Fuchs, S.M. (2012). Peptide microarrays to interrogate the "histone code". Methods Enzymol. *512,* 107-135.

Song, Q., Liu, G., Hu, S., Zhang, Y., Tao, Y., Han, Y., Zeng, H., Huang, W., Li, F., Chen, P.*, et al.* (2010). Novel autoimmune hepatitis-specific autoantigens identified using protein microarray technology. J. Proteome Res. *9,* 30-39.

Johnathan Neiswinger

jneiswi1@jhmi.edu

(724) 601-8303

Curriculum Vitae

Profile

Graduate student with passion for teaching and mentoring in the sciences. Experience in disciplines including organic and inorganic chemistry, biology, molecular biology, pharmacology, and bioinformatics.

Education

Johns Hopkins School of Medicine (August 2007-present)

Program: Pharmacology, PhD candidate

Honors:

- Appointed to teaching fellowship at Loyola University to teach Biology courses for three semesters
- Chosen to participate in newly formed course "Introduction to Effective Instruction"
- First Place Poster Session Award for exemplary research poster at Pharmacology research retreat

Geneva College, Beaver Falls PA (graduated May 2007)

Major: Chemistry (ACS-certified)

Minors: Biology, Math

GPA: 3.984 (4.0 scale)

Honors:

- Graduated Magna Cum Laude
- Member of Alpha Chi, a National Student Honor Society
- Recipient of Trustee Scholarship and Alumnus Chemistry Scholarship
- American Chemical Society Polymer Education Committee Award for Excellence in Organic Chemistry
- Presidential Scholarship for outstanding academic performance
- Dean's List every semester

Work History

Johns Hopkins University (Aug 2007-present)

- Graduate Student in the Department of Pharmacology studying kinase interactions with protein microarrays

The University of Akron (May 2006-Aug 2006)

- REU Internship studying elastic properties of virus films

Geneva College (Aug 2004- May 2005; Aug 2005-May 2006)

- Lab assistant for non-science major lab (incl. physics, biology, and chemistry)

Geneva College (May 2004 – May 2005)

- Paid musician (ambassador representing Geneva College nationally and internationally)

Research Interests

My doctoral work has focused on using protein microarrays to identify novel interactions between kinases and their respective substrates. Protein microarrays consist of thousands of proteins immobilized onto a solid surface in a grid-like pattern. They allow for the parallel experimentation of thousands of proteins at once in a single array the size of a microscope slide.

The first project is a collaboration with a bioinformatician, who predicted novel scaffolding proteins using protein-protein interaction databases. Treatment of a microarray with a kinase in the presence of ^{32}P-γ-ATP, and with and without a third "scaffold" protein, allows for the identification of kinase substrates that can only be phosphorylated in the presence of a third party protein (i.e., a scaffold). Using the hypothesized scaffolds on human proteome microarrays (which contain over 17,000 purified human proteins), we were able to identify a large number of proteins that met this condition.

The second project examines the apparent cross-talk between phosphorylation (by kinases) and glycosylation (by the enzyme OGT). O-linked glycosylation is a post-translational modification that occurs on S/T residues, similar to phosphorylation. Many

studies have shown that both phosphorylation and glycosylation can affect kinase activity and substrate specificity. Using antibody and click chemistry approaches that recognize glycosylated S/T, we were able to examine the glycosylation of over 100 kinases using a kinome microarray, which contains 350 human kinases.

Laboratory Skills

- High throughput purification of proteins, including active kinases
    - Oversaw the purification of the entire yeast genome (approximately 6,300 proteins)
- Microarray fabrication
    - Have created a "kinome" microarray containing 460 human kinases
- Radiation Safety Officer in lab (October 2011 – present)
    - No infractions in this time
- Tissue Culture
    - Maintained HEK293, HeLa, and NIH3T3 cell lines
- Proficient in western blotting, siRNA knockdown, *in vitro* enzyme assays, radioisotope experiments, DNA cloning, microarray analysis, and high throughput transformation and transfection.

Publications

1. **Neiswinger J**, Han Z, Zhu H, Hart GW. Probing the *O*-GlcNAcylation of the Kinome Using Protein Microarray Technology. 2014, in preparation.

2. Hu J, **Neiswinger J**, Zhu H, Qian J. Prediction of Scaffolding Proteins in Phosphorylation-Mediated Pathways. 2014, in preparation.

3. Wu XY, Renuse S, Zahari S, Nandini SA, Chaerkady R, Kim MS, Mohseni M, Zhong J, Yang J, **Neiswinger J**, Jeong JS, Newman R, Qian J, Zhu H, Vogelstein B, Park B, Pandey A. Activation of diverse signaling pathways by oncogenic PIC3CA mutations. ***Nature Communications*** 2014, in review.

4. **Neiswinger J**, Zhu H, Zhang J, Qian J. Protein Microarrays: A Versatile Tool for Scientific Discovery. ***Proteomics: Targeted Technologies, Innovations, and Applications,*** from Horizon Scientific Press's "Molecular Biology" series 2013, in review.

5. **Neiswinger J**, Uzoma I, Rho HS, Song G, Paul C, Jeong JS, Lu KY, Chen CS, Zhu H. Protein Microarrays: Flexible Tools for Scientific Innovation, ***Cold Spring Harbor Protocols: Budding Yeast*** 2013, in review.

6. Newman RH, Hu J, Rho H-S, Xie Z, Woodard C, **Neiswinger J**, Hwang W, Shirley M, Hu S, Cooper C, Jeong JS, Wu G, Lin J, Gao X, Ni Q, Dalby K, Ji H, Desiderio S, Birnbaum MJ, Cole PA, Knapp S, Ryazanov A, Zack DJ, Blackshaw S, Pawson T, Gingras A-C, Pandey A, Turk BE, Zhang J, Zhu H, Qian J. Construction of Human Activity-Based Phosphorylation Networks. ***Mol Syst Biol*** 2013; 9:655. doi: 10.1038/msb.2013.12.

7. Hu J, Rho HS, Newman RH, Hwang W, **Neiswinger J**, Zhu H, Zhang J, Qian J. Global analysis of phosphorylation networks in humans. ***Biochem Biophys Acta*** 2013 Mar 21. doi:pii: S1570-9639(13)00122-2.

8. Muratore KE, Seeliger MA, Wang Z, Fomina D, **Neiswinger J**, Havranek JJ, Baker D, Kuriyan J, Cole PA. Comparative analysis of mutant tyrosine kinase chemical rescue. ***Biochemistry*** 2009 Apr 21; 48(15):3378-86. doi:10.1021/bi900057g.

9. Hartschuh RD, Wargacki SP, Xiong H, **Neiswinger J**, Kisliuk A, Sihn S, Ward V, Vaia RA, Sokolov AP. How rigid are viruses? ***Phys Rev E Stat Nonlin Soft Matter Phys.*** 2008 Aug; 78(2 Pt 1):021907. Epub 2008 Aug 15.

Invited Talks

1. Probing the Cross-talk Between Glycosylation and Phosphorylation Using Protein Microarray Technology; Johns Hopkins University High Throughput Center Symposium, Baltimore, MD; November 15, 2012; Poster.
2. Identification of Novel and Isoform-Specific Akt Substrates Using Protein Micraoarray Technology; Johns Hopkins University Research in Progress, Baltimore, MD; November 28, 2011; Talk.
3. In Pursuit of a Kinome Network: Identification of Novel Kinase-Substrate Interactions Using Protein Microarray Technology; Geneva College, Beaver Falls, PA, November 12, 2010; Talk.

4. Identification of Novel and Isoform-Specific Akt Substrates Using Protein Microarray Technology; Johns Hopkins University Research in Progress, Baltimore, MD; February 15, 2010; Talk.
5. Using Protein Micorarrays to Identify Novel Targets of Kinases; Pharmacology Research Retreat, Baltimore, MD; September 29, 2009; Poster.